The Genetics and Spread of Amyotrophic Lateral Sclerosis

Jones, Ashley Richard

Awarding institution:
King's College London

THE GENETICS AND SPREAD OF AMYOTROPHIC LATERAL SCLEROSIS

Ashley Richard Jones

PhD in Clinical Neuroscience

Abstract

Our knowledge of the genetic contribution to Amyotrophic Lateral Sclerosis (ALS) is rapidly growing, and there is increasing research into how ALS spreads through the motor system and beyond. This thesis examines how genetic and non-genetic factors in ALS influence its spread.

The genetic methods employed were PCR, genotyping, AFLP, DNA sequencing and gene expression. The methods to examine spread were H&E staining, clinical history, age of onset (AOO), survival and health utility. Statistical procedures applied included regression analyses of genetic and non-genetic factors, maximum likelihood estimation of genetic-phenotype variance and health utility, RNA-sequence analyses, and differential gene expression analysis.

I found (A) that variation in the *ATXN2* gene contributes to ALS, as does variation in *C9ORF72* after correcting for the known *C9ORF72* pathological hexanucleotide repeat (HREM). (B) On comparing regions of the spinal cord, patterns of differential gene expression between ALS cases and controls appeared consistent with spread and pathology. Functional annotation clustering revealed these genes were mostly involved in blood vessel and angiogenin-like function, glycoprotein-based activity, and leukocyte activity. (C) A significant proportion of survival variance in ALS could be explained by genetic variance. There were SNPs that predicted survival and AOO, one showing epistasis with the *C9ORF72* HREM. (D) When modelling ALS progression using staging and a clinical trial dataset, the time and duration of each stage was statistically predictable. Staging also predicted health utility and other functional and psycho- metrics.

The spread and pathology in ALS spinal cord regions affected gene expression profiles, which is likely a consequence of genetic susceptibility in that region. Indicators of spread, AOO and survival, could be predicted using genotypes. Disease progression was predictable as measured by clinical staging and health metrics. In summary, ALS spread seems to occur at a fixed rate in an individual and is influenced by genetics.

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

My foremost acknowledgement, a word in which no way captures my appreciation for all his help, support, encouragement and friendship, is Professor Ammar Al-Chalabi. He has unfalteringly stuck by my side through these three years. He had faith in me, planted me in the deep end, and respected the evolution of my single-cell understanding of complex disease genetics into what is now, hopefully, in its first stages of embryogenesis, so that I may one day aim to be as good an academic as he is

Thank you Ammar.

I would like to acknowledge John Powell, whose profound grasp and conceptual understanding of our aims laid important foundations to how I approached several chapters in this thesis.

I would like to thank the members of Ammar's team; Kirsten, Ione, and Rubika, in which I worked with on several projects, and generally sat around discussing, exploring and despairing. To William and Aleksey who patiently helped me in the laboratory, especially when facing the temperamental contraptions. Note that Kirsten and William both exacerbated my coffee-addiction.

I would like to thank Dr Claire Troakes, coordinator of the MRC London Neurodegenerative Diseases Brain Bank, who never once grumbled despite my extensive tissue requests and my rummaging around clinical notes during very busy times. And Dr Andrew King who performed the H&E staining, which laid the foundation for disease spread in Chapter 6. Thank you to both.

I would like to thank Dr David To and Dr Caroline Johnston, bioinformaticians and system administrators of the NIHR Biomedical Research Centre for Mental Health. Specifically to Dr To for the RNA-seq help and Dr Johnston for help with imputation protocols. I'm very lucky to have such support almost any time of the day.

A big thank you to Dr Laura Ferraiuolo and Dr Janine Kirby from Sheffield University's Institute of Translation Neuroscience (SITraN), who helped me in genotyping and general lab techniques. Their warmth, support and openness to my lack of knowledge gave me faith that I could to grips with, what at the time was, very daunting lab techniques.

I would also like to acknowledge the Guy's Hospital Genomic Facility, especially Dr Efterpi Papouli, Dr Muddassar Mirza, and Dr Venu Pullabhatla. Their assistance in performing the gene expression analysis, when time was running out, was vital to this PhD.

I would like to formally thank the Medical Research Council who financed me and this PhD. I was very lucky and honoured to receive support from one of the top organisation in world for medical research. I hope I made the money go to good use.

I would also like to acknowledge the Motor Neurone Disease Association and European Network for the Cure of ALS. These two organisations are the cornerstones to anyone working in ALS today. Furthermore what they do is exceptional and the support offered to a young researcher like me was very warming.

A major thank you goes to my Father. Living in London isn't cheap and he did not once complain in supporting me. He encouraged me to live healthily, to travel to conferences around the world and to not work too hard. Note I achieved the travel advice only.

My final thank you goes to Gazal, and it is a big thank you! Gazal's love and patience with my insomniac dressing-gown work-ethic helped me with the freedom to push myself, while knowing I had that one person to fall back when I felt emotionally strained. She was always there, always supported me, and as a very talented clinical-scientist she contributed much to this thesis, more than she will recognise or take credit for. Thank you.

Dedication

I dedicate this thesis to the sufferers of amyotrophic lateral sclerosis. Even if only fractionally it brings us closer to a cure, the time spent on it would be of incalculable worth to me.

“All of old.

Nothing else ever.

Ever tried.

Ever failed.

No matter.

Try again.

Fail again.

Fail better.”

Samuel Beckett, Worstward Ho (1983)

Chapter 1 Introduction and Literature Review

1.1 Introduction

The cause and manifestation of Amyotrophic Lateral Sclerosis (ALS) is something we know little about, although in the last five years there have been accelerating advances in our understanding of the genetics, and interesting observations of how the disease seems to spread through the central nervous system (C.N.S.). This thesis examines both the genetics and the spread of ALS together, attempting to reconcile these two fundamental components of its pathology.

We are confident that in ALS there is an orchestra of different genetic mutations that contribute to this pathology. Some share common disease mechanisms; some we see more prevalent than others; some are much more severe than others; but ultimately we do not know the means by which any of these mutations cause ALS. Ways to explain these various findings include computational network-modelling of genetic-cellular pathways, high-resolution exome sequencing, and increasingly powerful genome-wide association tests involving thousands of cases and controls. The literature review below shows that 'older' methods have been the dominant means of ALS genetic discovery in the last few years, more so than the more recently devised methods that are listed above. There are many reasons these "new tools" have not yet reached their full potential in ALS research. It could be that these contemporary methods have yet to become useful for ALS in their development cycle, or that less contemporary methods are more useful for ALS because of the nature of the disease itself, which tend towards finding, in this case, more penetrant Mendelian mutations. Summating findings in western-dwelling Caucasians we find Mendelian mutations account for approximately 75% of familial ALS and 14% in those without a recognisable family history[1]. In addition we are beginning to see common functions among ALS genes, such as RNA processing[2] and protein recycling[3], which suggest they are of pathological importance in the disease.

ALS is increasingly being viewed as an oligogenic disease[4] and which genetic mutations are causal or modifiers of its phenotype are not yet clear. Why specific ALS genetic mutations correlate with a quicker disease progression is currently unknown; we hypothesise that perhaps

such genetic factors influence ALS progression by influencing disease spread through anatomy leading to a function of time. Nearly 150 years of clinical observations of ALS have informed research on the implicated neuroanatomy, but important questions are still left unanswered. For example, where does the disease begin (if anywhere), how does it spread in the C.N.S. and why is the disease highly variable in its duration and its presentation? Major obstacles stand in the way; for instance, symptomatological variability itself makes it difficult to study, the disease's wide pleiotropy with other neurodegenerative diseases is a confounding factor, and no convincing biomarkers are studiable as the patient is alive, and post-mortem proteomic and other biomarkers can be too profuse and too assorted to lead to clear hypothesis-generation.

The characterisation of ALS disease spread has been a relatively neglected area in ALS research. What is known is that the disease seems to begin at a focal location and moves towards an end-state, where ALS has engulfed the majority of upper and lower motor neural functionality. This is surprising because ALS is regarded as the result of a combination of genetic mutation and environmental exposure, and the idea that ALS is caused by ubiquitous aberrantly expressing genes and, at the same time, by an event at a specific anatomical location, which then progressively spreads throughout the motor system, is not immediately obvious. Clinical, neuropathological and neuroimaging studies show ALS focal-to-diffuse disease spread is evident; so then what is the contribution of the genetic mutations found in ALS?

The main experimental section of this thesis begins by testing genes suspected to be involved in ALS. This is mainly performed using genotyping and bioinformatic analyses. Following this I use gene expression to examine differentially expressing genes stratified by anatomy and inferred disease location and stage. I then use the genotype data to examine ALS phenotypes that are hypothesised to affect disease spread; age of onset (AOO), survival and onset location. Survival is an obvious indicator of disease spread and progression, but I believe that AOO and onset location is also, as they both directly influence ALS survival. And the final chapter examines the progression of ALS using King's ALS Clinical Staging, which standardises survival and onset location, to present a predictable model of disease progression measured using healthy utility, health-related quality of life and a psychometric measure of depression and anxiety.

My overarching aim is to present a round and wide exploration of the genetics and disease spread of ALS in the same conceptual framework.

1.2 Literature review of genetic discovery in ALS

1.2.1 Genetic Risk Factors

ALS is commonly considered a condition that is inherited in 5% and otherwise sporadic[5-7]. Although there are families in which transmission is clearly autosomal dominant with almost complete penetrance, in many cases the situation is less clear cut, with second or third degree relatives affected [8]. Because many do not know their relatives beyond such relationships, it is not clear whether a more distant family history is common or even relevant in ALS, but it is likely that at least some cases of sporadic ALS represent examples of familial clustering on a coarse scale that is not apparent to patients or medical staff. The risk to relatives in such families would therefore be greater than the background lifetime risk of 1 in 400 [9]. Thus the distinction between familial and sporadic ALS is to some extent artificial, although convenient for the purposes of research. There is now considerable evidence for a genetic contribution to familial and sporadic ALS. Several genes have been shown to be mutated in familial ALS, and these will be discussed here. Mutations of every such gene have also been identified in sporadic ALS, and while causation is more difficult to show in this situation, it seems reasonable to assume that a pathogenic mutation causing familial ALS is responsible for sporadic ALS when found in someone with no family history[10-12]. The heritability of sporadic ALS has been estimated as 0.61 (0.38 to 0.78) [13] and there are now several robustly replicated genetic associations. Although, in common with many other diseases, much of the heritability remains to be explained, newer genetic techniques such as next generation sequencing and epigenetics are likely to find the genetic variation responsible. Because the list of ALS genes is frequently changing, a complete list is available with relevant literature and bioinformatics links at the ALS Online Genetics Database, http://alsod.iop.kcl.ac.uk[14].

Several methods exist for gene identification. Family based methods generally rely on a statistical method known as linkage, although other techniques such as homozygosity mapping will work in specific circumstances. When family-based methods are not possible, association studies are usually the alternative, and these can either be based on study of a specific candidate gene or by an unbiased approach such as genome-wide association.

1.2.2 Genes identified through family-based studies

1.2.2.1 *SOD1*

Inherited forms of adult onset motor neuron degeneration were recognised in the 1800s [15], and ALS was one of the first conditions in which linkage to a gene was observed. Mutations in *SOD1*, responsible for about 2% of all ALS[10], were first identified in 1993[16, 17], and there are now more than 150 mutations of *SOD1* known affecting all parts of this small, five exon gene (http://alsod.iop.kcl.ac.uk)[18, 19]. The pathogenic mechanism remains unknown despite intensive efforts although it is clear there is a gain of some toxic function. One of the major problems to understand is how mutations occurring almost anywhere in the molecule[20] can result in the same toxic gain of function with the same mode of inheritance and the same specificity for motor neurons.

The commonest *SOD1* mutation in the US is A4V[21], accounting for nearly 50% of those with *SOD1* mutation[22] and associated with an aggressive form of ALS with short survival[23]. The alanine at codon 4 is essential for the structural integrity of the *SOD1* protein[24, 25], but this is not necessarily the mechanism of action as other mutations with no such role can result in ALS as well. Efforts to classify *SOD1* mutations by structural effect have not led to any obvious relationship between phenotype and mutation effect (http://alsod.iop.kcl.ac.uk).

The D90A mutation is the commonest worldwide and is interesting in several respects. First, it is the only mutation that shows a clear recessive inheritance[26]. There are individuals with ALS who

are heterozygous for D90A[12, 27], but it is likely that in these cases it is one of several contributors to the burden of disease liability. Second, the phenotype of those who are homozygous is characteristic, with slowly ascending symmetrical paresis. Third, the phenotype of those heterozygous for D90A is often more severe and aggressive than those with homozygous genotypes[12]. Fourth, because it reaches polymorphic frequency in parts of Scandinavia, the distinction between sporadic and familial disease is meaningless, as two carrier parents can occur by chance and the familiality of the disease is then a product of chance and sibship size, while in other cases, an affected homozygous individual having children with a carrier will generate an apparently dominant pattern of inheritance. Fifth, founder studies have shown a single worldwide founder based in Russia or Scandinavia about 50 generations ago[27]. In contrast to A4V, the D90A mutation does not appear to have much effect on any aspect of *SOD1* function or structure[26, 28].

The G93A mutation has been the subject of extensive research through transgenic animal models[29] but again, the mechanism of action is unclear but may include inducing apoptosis, mitochondrial dysfunction, and oxidative cellular damage[30-33].

1.2.2.2 *ANG*

Angiogenin (*ANG*) on chromosome 14q11.2 was originally screened as a candidate gene in Irish and Scottish families with ALS because it shares a metabolic pathway with the vascular endothelial growth factor gene, *VEGF*[34], which is implicated in ALS (see below). Furthermore, it is in linkage disequilibrium with the apurinic endonuclease gene, *APEX1*, which was associated with ALS in other studies[35]. ANG missense mutations have been identified in most Western European countries and the USA, in those with a positive family history and those with the sporadic form[36 37]. ANG mutations have also been found in patients other mutations, including FUS, TARDBP and SOD1, supporting an oligogenic bases for the cause of ALS [4, 38]

1.2.2.3 *TARDBP*

There remained no significant familial ALS genetic breakthrough for several years until a pathological clue led to the identification of the next familial ALS gene and a shift in our understanding of ALS. A major component of ALS pathology is the presence of ubiquitinated inclusions[42, 43]. Although it was known that neurofilamentous aggregates comprised at least some of these, the major component was unknown until discovered to be TAR-DNA binding protein 43 (TDP43), a protein involved in the response to HIV infection and in RNA splicing, with a major role in cystic fibrosis[44-48]. Wild-type (^{WT}) TDP43^{WT} is believed to be multifunctional, influencing RNA splicing inhibition and exon skipping, hnRNPs binding, transcriptional regulation, mRNA genesis, apoptosis, as well as protein-protein interactions, mainly through the C-terminus[44, 47, 49-51]. Linkage followed by sequencing of DNA from ALS families identified segregating mutations in the *TARDBP* gene on chromosome 1p36.22, coding for TDP43 [52], and functional studies supported a role for the protein in motor neuron death. Subsequently many other studies have replicated the original findings, and mutations have also been found in sporadic ALS[44, 45, 53]. The mutations nearly all affect the C-terminus; of the six ALS-causing mutations, all but one (D169G) are in exon 6, and all but one (truncating mutation Y374X) are missense mutations[54, 55]. The mutations result in redistribution of TDP43 from the nucleus to the cytoplasm in neurons and glia [49, 56] in the anterior horn of the spinal cord[57]. *TARDBP* mutations account for about 3% of familial and 1.5% of sporadic ALS[53].

1.2.2.4 *FUS*

Shortly after the identification of the pivotal role of TDP43 in ALS, linkage of other ALS families to chromosome 16q12 resulted in the next major target for gene identification[58]. Of the many genes in the target locus, one was similar to TDP43, and screening of this candidate, Fused in Sarcoma (*FUS*) identified many mutations in two pivotal studies, one from the UK and one from the US[59, 60]. Most mutations appear to be missense and based in exons 14 and 15[61]. These exons code for the C-terminus, with the commonest being R521C, which disrupts RNA binding and transcription processing, leading to a toxic-gain of function[59, 62, 63].

FUS protein is 526 amino acids long and structurally, functionally, and proteinopathically similar to TDP43[53]. Indeed, it has been considered that they might be acting in concert in ALS pathology[64], although current evidence points towards them being mutually exclusive. Inclusions immunoreactive for FUS have been located within nucleic and cytoplasmic regions in several studies[65, 66]. FUS^{WT} is multifunctional, and has been implicated in alternative splicing[67], genomic maintenance[68], and transcription factor regulation[69].

Interestingly, mutations of exons 3, 5 and 6 account for about a third of the mutations identified so far, and may be commoner in sporadic than familial ALS[70]. This part of the gene codes for the glycine rich N terminal and may result in aberrant mRNA synthesis[71]. *FUS* mutation is estimated to account for 4.4% of familial and 0.7% of sporadic ALS[59-62, 70, 72].

1.2.2.5 *OPTN*

Using homozygosity mapping in consanguineous families, a region of chromosome 10 was implicated in ALS in a Japanese study[73]. Mutation analysis of these and other samples from people with familial and sporadic ALS, revealed mutations in a gene known to cause primary open angle glaucoma, Optineurin (*OPTN*). Mutations have been found in both familial and sporadic ALS, one missense autosomal dominant (E478G), and two autosomal recessive, a deletion of exon 5 and a nonsense mutation on exon 12 (Q398X)[73]. In general, those with *OPTN* mutation had a slow progression of ALS. The mechanism through which the recessive and dominant mutations result in ALS may well be different, with a loss of function likely for the recessive, and a gain of function related to mislocalization of the OPTN protein possibly important for the dominant. The impact on OPTN protein of ALS and primary open angle glaucoma linked mutations is different, with different effects on localization and different effects on NFKb signalling, and this is likely to be the reason for the difference in phenotypic expression.

1.2.2.6 *DAO*

The D-amino oxidase gene, *DAO*, was found to be mutated in a family linked to chromosome 12q22-23. Functional studies support the pathogenic role of the mutation R199W, but other families have yet to be found harbouring mutations in this gene[74].

1.2.2.7 *VCP*

The valosin-containing protein gene, *VCP*, on chromosome 9 is involved in frontotemporal dementia with Paget's disease[75]. A study using whole exome sequencing in Italian families with ALS has identified mutation in *VCP* with supporting functional evidence, and it seems likely that such mutations will be confirmed as a rare additional cause of ALS[76].

1.2.2.8 *C9ORF72*

A GGGGCC hexanucleotide repeat expansion mutation on Chromosome 9 open reading frame 72 (*C9ORF72*) was found in 2011 through family linkage. The 9p21 locus was implicated via the same method in 2006[77, 78]and was replicated in several families over four years. Genome-wide association analyses (GWAS) confirmed this region's involvement in ALS in several western populations, narrowing it down to haplotype spanning three genes[79-81], Mps One Binder kinase activator-like 2B (*MOBKL2B*), interferon kappa (*IFNK*) and an open reading frame, *C9ORF72*. In 2011 a mutation linked with intron 1 of *C9ORF72* was found tagging this haplotype[82, 83].Although currently debated, it is believed the repeat mutation becomes pathogenic beyond 23 repeats, although it is mainly seen that expansion in ALS cases is much greater than this (so much so that it cannot be measured using regular PCR). Little is known how this causes ALS, however there are two dominant hypotheses, (a) nucleic RNA foci sequestration of RNA binding proteins and/or (b) cytoplasmic repeat-associated mutated proteins via non-ATG RAN translation.It's prevalence,

in a predominantly Caucasian population, is currently estimated to be 37.6% for familial ALS and 6.3% for without any obvious ALS inheritance[84]. This makes it is the most prevalent genetic mutation in ALS found to date. And evidence suggests that it significantly shortens survival (measured by duration) compared to more typical ALS presentation[85].

1.2.2.9 *UBQLN2*

A single point mutation on the Ubiquilin-2 (*UBQLN2*) gene, which encodes for an ubiquitin-like protein, was found X-linked dominantly inherited in a five generational family3. In the same study also confirmed four additional *UBQLN2* mutations in four other families showing no male-to-male inheritance. Consistent with X-linked inheritance AOO in males is much younger males because they are likely to be hemizygous for the mutation.

UBQLN2 mutations are especially interesting because Ubiquilin-2 skein-like protein inclusions are a common feature of sporadic and familial ALS and ALS-FTD post-mortem pathology, suggesting that they point to a common pathway in ALS pathology. This is likely to be linked to autophagy and protein recycling and shares functional similarity with the diffuse ubiquitin and p62 inclusions commonly seen in ALS post mortem tissue. Supporting the oligogenic nature of ALS, Ubiquilin-2 inclusions are seen in patients with known TDP-43, FUS, and ubiquitin ALS pathology.

1.2.3 Genes causing atypical ALS identified through family based studies

Mutations of several genes have been found to cause atypical ALS, including the vesicle-associated membrane protein-associated protein B (*VAPB*)[86], which usually causes a lower motor neuron predominant variant of ALS, alsin (*ALS2*), which usually causes a form of young onset hereditary spastic paraparesis[6, 87, 88], and senataxin (*SETX*), which causes a juvenile form of slowly progressive lower motor neuron ALS known as ALS4[89].

1.2.4 Genes identified through candidate gene association studies

1.2.4.1 *NFH*

Many genes have been proposed to be ALS susceptibility factors on the basis of association with ALS of common or rare variants within the gene. One of the first such replicated findings was of indel mutations in the neurofilament heavy gene, *NFH*[90] . Two studies found various deletions [90, 91] and another study an insertion[92], all in the *NFH* tail domain which contains a lysine-serine-proline repeat that is highly phosphorylated and involved in neurofilament assembly and spacing. All the indels were found in people with no family history of ALS, except one in someone with familial ALS[91], although segregation with disease could not be shown. Two other studies of familial ALS did not identify such variants[93, 94], but this is consistent with the idea that *NFH* tail domain indels increase the general liability to ALS without being of sufficiently large effect to cause disease. A large study of familial and sporadic ALS examining all three neurofilament subunits did not identify any association of variants with ALS, but the *NFH* tail could not be amplified and was therefore excluded from the analysis[95, 96].

1.2.4.2 *ATXN2*

Spinocerebellar ataxia type 2, caused by triplet repeat expansion in the ataxin 2 (*ATXN2*) gene is associated with motor neuron degeneration in some case reports[97, 98]. ATXN2 protein expression is elevated in ALS spinal cords and it interacts with TDP43 protein in a complex. Furthermore, TDP43 pathology is seen in spinocerebellar ataxia type 2, and ATXN2 is abnormally localized in ALS motor neurons[97]. Thus, the *ATXN2* gene makes an excellent candidate for involvement in ALS. Expansions of the triplet repeat in *ATXN2* are large in spinocerebellar ataxia type 2 (greater than 33 repeats), and normal in unaffected individuals (less than 27 repeats), whereas intermediate length polymorphisms increase the risk of ALS. The threshold at which repeats

become pathogenic are debated, and may be population-specific {Van Damme, 2011 #955}{Ross, 2011 #954}. CAA PolyQ interruptions have been associated with the ALS-related ATXN2 repeat length and not with SCA2, and therefore may be more important than the repeat length itself {Yu, 2011 #956}.

1.2.4.3 *SMN1 AND 2*

The survival motor neuron genes, *SMN1* and *SMN2* are involved in the commonest childhood onset motor neuron disease, spinal muscular atrophy[99]. Homozygous deletion of *SMN1* causes spinal muscular atrophy with a phenotype that can be ameliorated by increased copies of the *SMN2* gene which produces a less functional product[100]. Copy number variation (CNV) of *SMN1* appears to be a risk factor for ALS in some studies[101], although others do not find evidence for this[102].

Large-scale CNV analysis reveals a significant association of SMN1 duplications with ALS {Blauw, 2012 #1452}. Retrospective meta-analyses of ALS genome-wide copy number variation association studies help confirm this finding {Blauw, 2010 #431}{Cronin, 2008 #432}{Wain, 2009 #814}. Copy number studies are particularly sensitive to biases caused by quantitative PCR, prone to false-positive results by differential errors {Barnes, 2008 #1453}, which could explain the multiple null findings surround the SMN genes in ALS. This cause of error could be particularly confounded in the analyses of SMN as there share high sequence homology.

1.2.4.4 *VEGF*

The vascular endothelial growth factor gene, *VEGF*, was identified as an ALS gene serendipitously. VEGF is involved in the vascular response to hypoxia and is therefore important in ischaemic heart disease[106-108]. A mouse model of cardiovascular disease using knockout of the hypoxia response element of the *VEGF* gene had a phenotype with striking similarity to ALS[106] and a subsequent association study of the equivalent human single nucleotide polymorphisms (SNPs) found association in three of four populations studied, as well as in a joint analysis[109]. These SNPs have not been associated with ALS in genome-wide association studies or replication studies and the original finding in humans may therefore have been false positive[110, 111], although a second study suggested that the effect may be sex dependent[112, 113]. In any case, the mouse model provides evidence of the importance of VEGF to motor neurons and the rationale for generating a VEGF-based treatment.

1.2.5 Genes identified through genome-wide association studies of unrelated samples

1.2.5.1 *DPP6*

There have been several independent genome-wide association studies in ALS, several of which have used a second cohort for replication of top hits[79-81, 114-117]. The first replicated gene variant implicated through such studies and achieving genome-wide significance was of the dipeptidyl peptidase 6 gene, *DPP6,* identified through a two stage study of Dutch, Swedish, Belgian and US samples[116]. The same variant was the top hit in an Irish study but did not achieve genome-wide significance[114]. Although this finding has been confirmed in an Italian study [118], subsequent joint analyses have not confirmed this association, but this may be a product of the winner's curse phenomenon[119], where an initial finding shows a higher effect size than subsequent studies. Consistent with this idea, a weak signal not achieving genome-wide significance has been seen in association studies from samples of Italian[120, 121], Polish[117], and Chinese[122] populations.

1.2.5.2 *ITPR2*

Similarly, the *ITPR2* gene, coding for inositol 1,4,5-triphosphate receptor type 2, was associated with ALS in a Dutch population[115] and the finding confirmed within the same study on an independent population from Sweden, Belgium and Holland . Again, subsequent studies have not confirmed this[121, 123].

1.2.5.3 *UNC13A*

A large study using samples from many groups and again, with an internal replication sample, identified variation in the unc-13 homolog A (C. Elegans) gene, *UNC13A* as strongly associated with ALS[79]. This finding showed strong association in a joint analysis with samples from a third independent data set[80] and has been replicated in an East Asian study {Iida, 2011 #1454}. The *UNC13A* gene is interesting because of the role of unc13 proteins in neurotransmitter release and influences on glutamatergic transmission particularly[124].

1.2.6 Genes identified through other methods

1.2.6.1 *ELP3*

The elongation protein 3 gene, *ELP3*, is part of a complex involved in RNA processing through elongation of messenger RNA, histone acetylation and modification of tRNA wobble nucleosides[125-128]. A study using microsatellites to perform genome-wide association in three populations identified variation in intron 10 of the *ELP3* gene associated with ALS[129]. An independent mutagenesis study of *Drosophila* studying genes involved in axonal guidance and

development, identified mutations of *ELP3* as a cause of defective neuronal guidance and function. Knockdown of *ELP3* in zebrafish resulted in defects of motor neuronal growth and branching. *ELP3* is a good candidate as an ALS gene because, like TDP43, FUS, SMN and *ANG*, it is involved in RNA processing, and defects of *ELP1* and *ELP4* cause familial dysautonomia and Rolandic epilepsy, two other neurodegenerative diseases[130, 131]. Whether ELP3 is a causative gene or a disease modifier is yet unknown.

1.2.7 Genes modifying Age of Onset

1.2.7.1 Using linkage analysis

1.2.7.1.1 *SOD1*

Homozygosity for *SOD1* mutations is known to cause earlier age of onset, although G114A G37R, I104F, and L38V *SOD1* mutations have been associated with younger ALS onset[132]. There is also evidence that ALS age of onset (AOO) is approximately 29% hereditable for families with A4V and D90A *SOD1* mutations[132], but this is yet to be replicated.

1.2.7.1.2 *APOE*

Two studies have found an association between *APOE* genotype and age of onset of ALS[133, 134]. This was also supported by a transmission disequilibrium test using 508 families, which found APOE-2 to be protective in ALS leading to later AOO.

1.2.7.2 *Using association analyses*

1.2.7.2.1 *CHGB*

This gene was chosen as a candidate due to its functional relationship with *SOD1* and it was found that the P413L mutation was associated with a younger ALS onset by up to almost a decade[135].

1.2.7.2.2 *TXNRD1*

TXNRD1 was chosen due to its redox functionality in ALS, again relative to the oxidative dismutase function of *SOD1*[136]. An intronic SNP rs6539137 of this gene was associated with earlier AOO, but only in males.

1.2.7.2.3 *1p34.1*

A locus modifying AOO, reducing it by up to over 2 years, was identified at 1p34.1 by an international association meta-analysis[137]. The two SNPs identified were rs3011225 and rs803675, in a population of 4243 cases and 5112 European controls.

1.2.8 Genes modifying survival

1.2.8.1 *KIFAP3*

A genome-wide study of ALS survival found variation in the kinesin-associated protein 3 (*KIFAP3*) gene associated with a 14 month survival advantage[138]. The study used clinic-based samples and a subsequent smaller study using population-based samples did not find evidence for the association[139]. However, clinic populations exclude short survivors and it may be that the population-based study was therefore underpowered.

1.2.8.2 *UNC13A*

SNP rs12608932 in intron 20 of the UNC13A gene was associated with survival[140]. rs12608932 was then replicated in follow-up multi European population of 1767 cases and 1817 control within the same study. UNC13A function is implicated in synaptic neurotransmitter release at neuromuscular junctions.

1.2.8.3 EPHA4

In 2012, EPHA4 was identified as a modifier of survival in a Zebrafish model{Van Hoecke, 2012 #1455}. Genetic inhibition of EPHA4 rescues $SOD1^{MT}$ SOD1 Zebrafish, and increases survival in two murine models. In humans, increased EPHA4 expression worsens AOO and survival, with loss-of-function associating with increasing survival time. The proposed pathological mechanism is an EPHA4 gain-of-function which contributes to the axonopathy prevalent in ALS and modifies motor neuronal vulnerability to the disease.

1.2.9 Genes modifying onset site

Bulbar onset is significantly more frequent in ALS cases with FUS and TARDBP mutations[72, 141], with FUS generally associating with a shorter survival than TARDBP. It is known that bulbar onset in men is also more frequent[9, 142]. Overall no direct genetic modifier of onset site has yet been found.

1.2.10 Genes modifying sex ratio

The prevalence and incidence of ALS is more in men by approximately 1.3[143], although this ratio is dependent on AOO[144]. Furthermore sex influences site of onset with an increased incidence of bulbar-onset ALS in women and flail-arm presentation in men.

1.2.10.1 Using linkage analysis

1.2.10.1.1 *SOD1*

There is evidence that genes influence sex dependent penetrance in a four-generational family with an autosomal dominant *SOD1* mutation (Phe20Cys) which reduced penetrance in the women[145]. *SOD1* influence on sex differences is supported by mouse models, where disease initiates earlier in males with the G93A mutation, but an alternative explanation may be the lack of protective properties from the heightened oestrogen in females[146].

1.2.10.2 Using association analysis

1.2.10.2.1 *VEGF*

A VEGF SNP allele from a haplotype associated with ALS has also been associated with increased incidence of ALS in women in Germany[112].

1.2.10.2.2 *TTC7A*

An intron SNP (rs735888) allele found at TTC7A on Chromosome 2 associated with ALS and is found significantly more in Caucasian women for sporadic cases. This SNP did not associate with ALS for men exclusively[147].

1.2.11 Summary

Genetic studies of ALS are beginning to shed light on the underlying pathways that lead to motor neuron degeneration. Axonal degeneration, synaptic signalling and RNA processing are themes that are emerging. For example, TDP43, FUS, *ANG*, SMN and *ELP3* are all RNA processing proteins. After a decade of effort at understanding ALS with models based on just one gene, *SOD1*, we now have many genes to build a picture of the pathogenic process with several more on the way. The relationship between ALS and frontotemporal dementia is becoming clearer as genes common and genes distinct for the two diseases become apparent, and with our greater understanding comes the hope that a new treatment will be generated.

ALS presentation is heterogeneous and it is unsurprising that elucidating the genetic and characteristics of its progression over time has proven challenging. There are significant predictors of ALS survival using genetics that translate directly into either faster disease progression or indirectly through causing bulbar-onset, which will cause respiratory death quicker due to its early anatomical involvement, or causing earlier AOO which will prolong ALS duration. Outside of genetics, these factors are known predictors of ALS disease progression. And although the genetic basis of these is still largely unknown, research is advancing in characterising how the disease actually progresses over its course. This has introduced the idea of health states, referred to as ALS stages. It has been found that the disease stages of ALS can be predicted in populations, despite its heterogeneity, informing us deeply about how the disease progresses and also giving us scope to provide better healthcare.

1.3 Literature review of disease spread and gene expression

1.3.1 Disease spread

The concept of how ALS spreads to degenerate motor neurons can be described by “dying-back” and “dying-forward”[148, 149] axonal pathology. It is possible that either upper motor neurons (UMN), lower motor neurons (LMN), or neuromuscular junctions, spread ALS to each other in a specific direction. Dying-back refers to how cell-death begins in the cell body and works its way “back” to through the corticospinal motor axon (possibly from UMN to neuromuscular junctions). Dying-forward refers to how the disease may begin at an axonal location (or perhaps junctions) and works itself towards the cell body. Much evidence points to UMN and LMN degeneration occurring separately in terms of progression and spread, but both UMN and LMN symptoms seem to share the same onset location, suggesting some kind of relationship between the two motor neuron types[150, 151].

1.3.1.1 Imaging

A reason why understanding disease spread in ALS has been so problematic is the disease situation and technology. ALS has a rapid progression rate with patients typically dying within three years of diagnosis. This does not leave much time for research, which in many cases can be regarded by the patient as undesirable due to the futility of the prognosis. The use of electromyography and motor unit number estimation (MUNE) techniques in the spinal cord, and functional neuroligand imaging and repetitive transcranial magnetic stimulation (rTMS) techniques in the motor cortex have had limited success. Comparing methods and qualitatively different neuroanatomical structures is also difficult. Spatial and temporal resolution is a problem in characterising cellular-level pathological changes into macro-level structural transformations. Although this may be slowly being overcome by technologies using advanced pixel, structural and diffusion tensor techniques[152], and imaging genomics. And, as in genetic research, disease heterogeneity impedes a clear understanding of how ALS spreads.

The extent of connectivity in the motor network has been described as a correlate of survival, suggesting that disease spreads through functional connections and this progression is caused by inter-neural transmission of a disease agent[153]. Another study has shown structural connectivity to become pathological at later disease stages as the disease goes on to progress to frontal and parietal regions[154], further supporting a more probable rostral-caudal spread.

1.3.1.2 Pathology

Neuropathological examination of post-mortem brain and spinal cords can be another approach to learn about ALS spread. It is believed that disease onset location corresponds to motor neuron loss[151] and that ALS directionality of spread is more frequently outward in a rostral-caudal direction for LMNs and a medial-lateral direction for UMNs, although findings have been mixed[155]. Pathological studies do show that motor neuron loss decreases gradually as you move away from the disease's neuroanatomical epicentre[151], and that the direction of spread influences survival[155].

These findings correspond with the clinical features of the disease, with greater frequency for spreading ipsilaterally for LMN symptoms and contralaterally for UMN symptoms[156], and with greater frequency moving bulbar to spinal rather than the converse[150].

The mechanism of pathology has been hypothesised to be caused by an initial, possibly stochastic[156], insult at a motor corticospinal location where there is genetic susceptibility. This leads to misfolded proteins, a known pathological consequence of some genetic mutations in ALS[157, 158]. Spread can then occur like a prion disease, with misfolded proteins aggregating, then binding and sequestering functional proteins as degeneration moves into a new motor cell. In genetic and neuropathological research, there are multiple dysfunctional pathways in ALS that may facilitate this[159] as protein aggregates are transmitted across cells[160].

1.3.1.3 Clinical symptomatology

In most ALS patients, the disease seems to begin in one limb[150] and over time become diffuse through the corticospinal motor neurons causing a heterogeneous clinical presentation[156]. Early studies of ALS found that it spreads contiguously more often and quickly, that this spread is uniform in its progression, and it was hypothesised that this was a positive sign of propagation[161, 162]. The King's clinical staging system, validated in Chapter 5, helps confirm the uniformity of disease progression nearer disease onset and greater clinical heterogeneity towards the end of the disease[163].

The direction of spread seems to influence survival, when comparing cases with a uni-focalonset which spreads caudal-rostrally, compared to cases with disease spreading or leaping to further anatomical regions[155].

1.3.2 Disease-spread using health states

Phenotypes such as AOO, bulbar or limb onset, and sex can help estimate the overall progression (or survival) of ALS, but do not inform us about how the disease progresses. Studying ALS progression through time has indicated that symptomatological accrual, statistically, is curvilinear[164], in which you can use early symptom and diagnostic information to predict ALS prognosis[165, 166] and survival[167]. However, the evidence also shows that progression of ALS in one individual may be linear but varies significantly between patients[162, 168].

The utility of rendering ALS disease progression predictable is vast. It provides a categorical context of which stage of ALS a patient is in. This can facilitate evaluative estimations in terms cost-effectiveness and efficiency of treatments, by for example using quality of life scores. Similarly, the health care needs of an individual can be addressed more effectively as we become more informed. It can also help facilitate clinical trial design and analyses, standardise trial data for better use in statistical methods, and similarly provide a framework in ALS longitudinal research.

Although several staging systems for ALS have been suggested in the past, the first formal staging system that used health states to map ALS clinical milestones was published in 2012[163]. This found that ALS disease progression could be standardised across patients and that time to health states was predictable. The difference between ALS staging and ALS-FRS is that the former captures information about clinical-based pathological progression, whereas the ALS-FRS captures information regarding progressive motor function. Motor function can oscillate over time whereas pathology is made more uniform, categorical and progressive, using the staging system.

1.3.3 Gene expression analyses

1.3.3.1 Mouse models

Mouse models are beneficial as you can measure disease spread across time across anatomy. However, the main focus has been temporal capture of gene expression profiles in lumbar spinal cords. Gene expression profiles, typically taken at 3-4 different time-points throughout the disease's duration, vary radically.

However, there is some replication between murine studies and between time-points[169], especially for vimentin (Vim)[170-176] and the cathespin gene family[170, 172-178]. Furthermore genes Cathepsin B and D, GFAP and SERPINA3 have been shown to differentially express in two or more mouse studies and two or more human studies[179-182], bridging the gap across species. Gene ontological analyses of replicated ALS mouse genes involve immune response, lysosome, metal ion binding and mitochondrion function to be either significantly under or over represented[169]. Whether these functions are disease responses is unknown. For example immune, mitochondrial and lysosome responses could be due to the motor cell failing or being under attack by some conformationally changed protein. They may not be specific enough to provide convincing pathways of ALS pathology. Metal ion binding dysfunction may be specific to the *SOD1* mutations studied in the murine cohort.

How applicable mouse pathology is to human pathology in ALS is debatable, as most have, up to now, been based on the *SOD1* mutation model. *SOD1* is the third most prevalent genetic mutation known in ALS today, but can present with a different phenotype compared to TARDBP and *C9ORF72* mutations carriers[183]. On the other hand, keeping the mutation constant may be beneficial as there is otherwise variance across microarray platforms in addition to variance caused by genetic-driven heterogeneity. Homogeneity is harder to achieve in humans.

1.3.3.2 Human C.N.S. samples

Similarly to mouse models, the majority of human gene expression projects have focused on lumbar spinal tissue, although cortical, cervical, ventral horn, sensory cortical and grey matter samples have been analysed. The majority of ALS cases have been sporadic. On average most papers reporting human C.N.S. gene expression studies have analysed about eight cases and seven controls. Since 2005 the introduction of laser capture microdissection of motor neurons in the spinal cord has been a major focus, although mixed-cell samples are still used.

There is correspondence between the functions implicated in motor cortical and spinal cord studies; these involve the cytoskeleton, protein turnover, apoptosis, and neurotransmission [180, 182, 184, 185 186]. Cytoskeleton defects have long been attributed to ALS pathology[187] and have been hypothesised to interact with ALS gene mutations such as those in *SOD1*[188] and *VAPB*[189]. Protein turnover is believed to be a central problem in ALS pathology, and this has been confirmed by the recent discovery of mutation in *UBQLN2*. This protein associates with ubiquilin and p62 proteins[183, 190]; two problem proteins consistently found to differentially expression throughout post-mortem corticospinal tracts. Apoptosis also has a long history in ALS research, and has been studied in relation to oxidative stress caused by *SOD1* mutations[191]. Neurotransmitter regulation has not been majorly implicated in the pathology of ALS before.

In terms of disease progression, a Japanese group[192] were the first to compare spinal motor cells (via LCM) and ventral horn tissue samples, finding a different gene expression profile for the latter. Examination of exon splicing between 12 cases and 10 controls in lumbar samples confirmed clinical and pathological evidence that the disease probabilistically progresses caudally[193]. They also compared motor neurons to adjacent anterior horn samples, examining lateral spread, and found differential gene expression and aberrant exon splicing is more evident in motor neurons, and that there is a greater presence of aberrant exon splicing in comparison to differential expression[193]. This supports the hypothesis that ALS is largely due to a malfunction in RNA processing functionality (see Section 1 of this review).

Problems exist in human C.N.S. gene expression studies due to the wide variation of cases, associated (and often unknown) genetic mutations, array platforms, sample sizes, tissue (and blood) samples, bioinformatic and statistical approaches, set differential expression and cluster thresholds, concomitant medication use status, and underlying clinical heterogeneity. Furthermore few studies have taken an internal control in their cases, controlling person-specific gene expression variation with gene expression caused by ALS.

1.3.3.3 Human periphery samples

Extra-C.N.S. periphery samples have been taken from lymphocytes[194] and mononuclear cells[195] from venous blood, and myocytes from muscle biopsies[196, 197]. There is correspondence in differential expression and functional enrichment analyses between blood samples and the C.N.S. findings described above, in genes involved in protein processing, RNA post-transcriptional modification, and inflammation[194, 198]. Furthermore, differential expression of TARDBP was identified in lymphocyte samples[194], consistent with known ALS genetic mutations in this gene. There is poorer correspondence between biopsy samples and C.N.S. samples[199].

1.3.4 eQTL analyses

It has been hypothesised that SNPs associated to traits are likely to be expression Quantitative Trait Loci (eQTL)[200], genomic variants that affect the expression of a gene either nearby (cis-acting) or at a genomic distance (trans-acting). This is primarily because greater variation of a complex can be captured using eQTLs than just the underlying SNP. However, eQTL have been successfully implicated in other complex diseases. The first and only eQTL analysis in ALS used a genome-wide marker dataset with gene expression levels from peripheral blood[201]. Eight SNP-transcript pairs were found and replicated to significantly modulate CYP27A1 expression.

CYP27A1 mutations cause cerebrotendinous xanthomatosis, a disease of upper motor neurons that can mimic primary lateral sclerosis[202, 203], making it a likely pleiotropic candidate in ALS.

1.3.5 Summary

The question of how ALS is both caused by ubiquitous genetic mal-expression and demonstrates focal-to-diffuse spread can be greater informed by corroborating the above methods. For example, we can regard the oligogenic combinations of ALS mutations as either causing a motor cell susceptibility or producing toxic aggressors, that facilitate some stochastic or non-genetic event to ultimately spread through functional and structural networks, Evidence predicts that this spread will be more often moving rostral-caudally and that the extent of its diffusion, or multi-distal disease locations, affects the heterogeneity and severity of the disease.

Few studies have examined disease spread in relation to differences in gene expression. However, one confirmed that there is a preference for ALS to spread in a caudal direction. Findings highlighted in gene expression research implicate common functions found throughout motor cortical, spinal and peripheral samples, perhaps suggestive that these functions are important in the wider spread of ALS in the body. Many however can be attributed as tissue-specific or cell responses to a pervasive neurodegenerative disease. Interestingly, one study found a greater involvement of aberrant splicing compared to differential expression. eQTL studies in ALS have yet to begin as a primary method, but the pioneering study identified a very plausible candidate gene, CYP27A1, known in a disease with similar dysfunction of the motor cortex.

Chapter 2 Genetic analyses using genotypes: *ATXN2* and *PLCD1*

2.1 Introduction

The genetic work in this and the following chapter builds upon previous findings from collaborators or publications. The genes examined are Ataxin2 (*ATXN2*) and Phospholipase C, Delta 1 (*PLCD1*) which are detailed in the literature review.

The aim of this work was to establish a better genetic understanding of whether or how these genes are involved in ALS. To do this I employed mostly bioinformatic methods using existing data. I will describe briefly describe the lack of understanding concerning how these genes contribute to ALS and the methods I took to resolve this.

2.1.1 *ATXN2*

In exon 1 of the *ATXN2* gene, CAG-trinucleotide expansion repeats greater than 33 were discovered to cause Spinocerebellar Ataxia 2 (*SCA2*). In 2010 it was also found that between 27 and 33 associated with ALS [97]. The actual repeat size, sequence, the mutations' exclusivity to ALS, the extent of causation and risk[204], and the intriguing idea that variable repeat lengths can cause fundamentally different diseases have been debated[205-207].

Notwithstanding this debate, the evidence that intermediate *ATXN2* polyQ repeats cause ALS now has several sources; by association of intermediate repeats in several populations[97, 204, 205, 207-213], the ataxin-2 protein interaction with TDP-43[97], and the *SCA2* Ataxia-ALS symptomatological overlap.

Lahut and Omur et al. (2013) found a 15-SNP haplotype in Turkish cases associated with ALS, which spanned *ATXN2* and its neighbouring gene *SH2B3*[213]. They were unable to find a single SNP showing significant association with ALS, but the haplotype's odds ratio (OR) was calculated at 2.23.

In collaboration with this research group I aimed to replicate this finding using imputation and association analyses of genotypes and haplotypes between ALS cases and non-ALS controls across four international populations. I believe that exploring the ALS-associated haplotypic background of *ATXN2* would help clarify whether intermediate *ATXN2* repeats cause ALS. My approach is under the hypothesis that specific haplotypes cause risk of pathological expansion, similiar to the proposed scenario found with the *C9ORF72* gene and the HREM.

2.1.2 *PLCD1*

Statistically ALS-associated markers from a microsatellite Genome Wide Association Study (GWAS) implicated the *PLCD1* gene in ALS[214]. In the same study, fine-mapping the alleles in the *PLCD1* marker (D3S1298) were found to associate with UK and Belgium ALS populations, but not USA. In a recent study[215] PLCδ1 (the protein of *PLCD1*) and *PLCD1* gene expression was shown to be up-regulated in ALS *SOD1*^{G93A} mice. It is also a good candidate gene for ALS as it is involved in excitotoxicity function, a putative pathological pathway in ALS. In collaboration with this study[215] I imputed genotype markers around *PLCD1*, and ran association analyses to ascertain a functional or mutation-tagging variant, and an ALS-related haplotype.

2.2 Methods

2.2.1 Sample population

All samples used in this chapter were from people with apparently sporadic ALS. The UK DNA samples are derived from blood and were given by the UK National DNA Bank for Motor Neuron Disease Research (MNDA DNA bank). The MNDA DNA bank acts, among other things, as a depository for DNA taken from ALS patients from 20 UK hospitals. UK control samples were taken from the Depression Case Control (DeCC) study[216], the Bipolar Affective Case Control Study (BACCS)[217], data deposited by Panos Deloukas from the Wellcome Trust Sanger Institute (Cambridge, UK) and published online from the British 1958 birth cohort DNA collection.

In addition to UK samples, sporadic ALS cases and control samples were used from Ireland, The Netherlands and USA populations. The details, including demographics, methods of collection, and quality control procedures can be found Cronin et al. (2008)[114], Van Es et al. (2009)[79] and Landers et al. (2009)[218]. The breakdown by population is shown in Table 2-1.

Country (N)	***Cases***	***Controls***	***% Female***	***Genotype Markers***
Ireland (432)	221	211	47%	561466
Netherlands (2113)	1046	1067	41%	535468
USA (1533)	745	808	38%	307790
UK	663	4519	51%	584414

Table 2-1. Characteristics of international cases and controls before quality control

All participants were of white European ancestry and were matched for geo-ethnic origin. All participants gave informed written consent, and the research ethic committees from each institution gave approval for this study.

2.2.2 UK Sample Collection

All UK DNA samples were previously extracted from blood samples by the various institutions involved. DNA was extracted within one week of bleeding and stored at the UK DNA Banking Network in Manchester. Samples were bar-coded in a tracking system to reduce risk of clerical error.

Blood samples from consenting ALS participants of the MND DNA Bank were used only. For sample homogeneity, selection was based on no family history of ALS, being of white European ancestry, and disease onset beginning January 2002 or after.

UK Control samples were obtained from the Depression Case Control (DeCC) study, the Bipolar Affective Case Control Study (BACCS), from Panos Deloukas of the Wellcome Trust Sanger Institute (Cambridge, UK) and the British 1958 birth cohort DNA collection. This project was ethically approved at all institutes involved; see Shatunov et al. (2010)[80]. For Ireland sample collections methods please see Cronin et al. (2008)[114], for Netherlands please see Van Es et al. (2009)[219], and for USA please see Landers et al. (2009)[218]

2.2.3 UK Genotyping

Please see Cronin et al. (2008)[114], Van Es et al. (2009)[79] and Landers et al. (2009)[218] for non-UK genotyping methods. Genotyping UK samples were processed using HumanHap550 BeadChips (Illumina, CA, USA) at UCL Genomics, London, UK. The raw data were analysed, quality checked and genotypes called using BeadStudio (Illumina). The output allows conversion into PLINK file formats for association analysis.

2.2.4 Genotype Statistical Quality Control

Raw genotype images files were converted to a Pedigree file (PED) using Illumina BeadStudio.

Statistical quality control excluded individuals with sex status incongruent with genetic sex, those marked for exclusion by the genotyping institution, those with genotyping < 0.98 per individual, those with phenotype missingness < 0.98, those with low heterozygosity ($p < 0.05$), those exhibiting identity by descent sharing >5% of alleles with another participant, and those not of white European ancestry by analysis of ancestral markers..

Excluded SNP markers were those with a minor allele frequency less than 0.015, SNP missingness <0.98, significant departure from Hardy-Weinberg equilibrium ($p < 0.001$), those marked for exclusion by the genotyping institution, and those on chromosome X (see Script A1 in Appendices).

2.2.5 Phasing and imputation

Phasing and imputation was completed using 1000 Genomes v2.20101123 autosomal release (http://www.1000genomes.org/), performed using MaCH 1.0.18

(http://www.sph.umich.edu/csg/abecasis/MACH/index.html) and minimac (http://genome.sph.umich.edu/wiki/Minimac). I performed association analysis on imputed data using ProbABEL (http://www.genabel.org/packages/ProbABEL), mach2dat (http://genome.sph.umich.edu/wiki/Mach2dat:_Association_with_MACH_output), and PLINK v1.07[220], after creating input files using the GenGen (http://www.openbioinformatics.org/gengen/) conversion tool. For GenGen, I set the r^2 threshold as 0.3 and, based on genotype posterior probabilities derived in MaCH, a quality control threshold of 0.9 (see Script A2 in Appendices).

2.2.6 Association and haplotype analyses

I used logistic regression, modelling case-control comparisons with sex as a covariate, and with population stratification covariates derived from SmartPCA and EIGENSTRAT[221] (see Script A3 in Appendices), in PLINK[220] (see Script A4 in Appendices).

I set the Bonferroni-corrected threshold of statistical significance at 1.13×10^{-7}, and the genome-wide threshold as $p = 5 \times 10^{-8}$. I used the R statistical Bioconductor package *ggplot2{Wickham, 2009 #1470}* to construct Manhattan and Q-Q Plots, LocusZoom (http://csg.sph.umich.edu/locuszoom/) for regional Manhattan plots, and Haploview for haplotype block maps (http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview). To correct for genomic inflation, *p* values were adjusted by the median χ^2 statistic (genomic control factor). I ran all genetic tests using additive models.

For haplotype analyses I used PLINK v1.07 sliding windows between 2 and 100 SNPs to explore significantly associated haplotypes using the phased imputed dataset (see Script A5 in Appendices). The prevalence of the most associated haplotypes were refined in Haploview and tested for their association with ALS using an omnibus haplotype association analysis.

2.3 Results

2.3.1 *ATXN2*

Using the 1000 genomes reference panel I imputed genotypes in and around *ATXN2*, including neighbouring genes to include *SH2B3*. Sliding windows analyses identified significant haplotypes associated with ALS in both UK and international datasets. The most significant associations were variations of a 47-SNP haplotype between rs10849944 and rs1544396, which incorporates *SH2B3* and *ATXN2* and the originally reported 15-SNP haplotype[213] (see Appendix Table A1). The frequencies of these occur approximately 45-50% in cases and 50-55% in controls, and none of these were significant after correcting for Bonferroni multiple testing.

This analysis also identified Lahut and Omur et al.'s (2013)[213] haplotype as the second most significantly associated with ALS in the international and UK datasets, but was not significant after multiple-testing correction.

It is likely that this correction is too stringent as there seemed to be distinct associations in multiple populations. I analysed the original 15-SNP haplotype as reported by Lahut and Omur et al. (2013)[213], using a conditional haplotype-based association test with a generalised linear model (GLM). Two SNPs were missing from out dataset, which reduced it to a 13-SNP haplotype.

Using a GLM meant we could account for population stratification using principal components. I found a significant association of the 13-SNP haplotype with ALS in the international population (OR: 0.48, p = 0.002). When I examined populations individually I found a significant association with the UK (OR: 0.49, $p = 1.97 \times 10^{-4}$) and near significant association with the Netherlands (OR: 0.47, p = 0.06) (see Figure 2-1 for 13-SNP haplotype). For the USA and Ireland this was not significant.

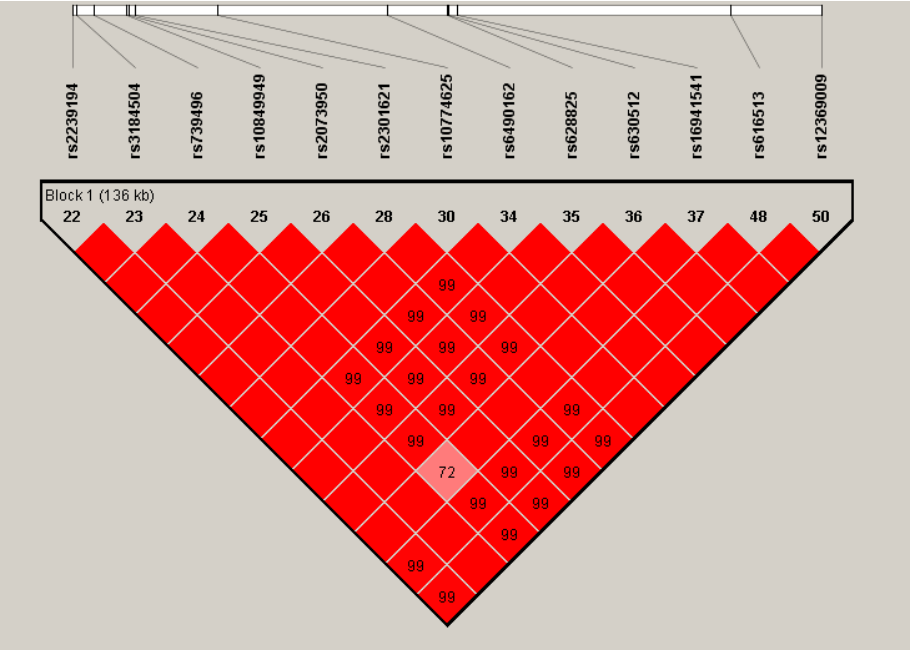


Figure 2-1. *ATXN2-SH2B3* haplotype associated with UK ALS cases.

2.3.2 *PLCD1*

Using the 1000 Genomes reference panels I imputed the region around *PLCD1* in both UK (cases n = 599, controls n = 4142) (see Figure 2-2) and international datasets (cases n = 2611, controls n = 6228). Only one genotyped marker was found in *PLCD1* and I found no association in *PLCD1* or in neighbouring SNPs before or after imputation (see appendix Figure A1 for international dataset Manhattan plot). I began with 28 genotype markers in and around *PLCD1* and imputed 1774, covering 11 genes in the neighbouring area from *ITGA9* to *OXSR1*.

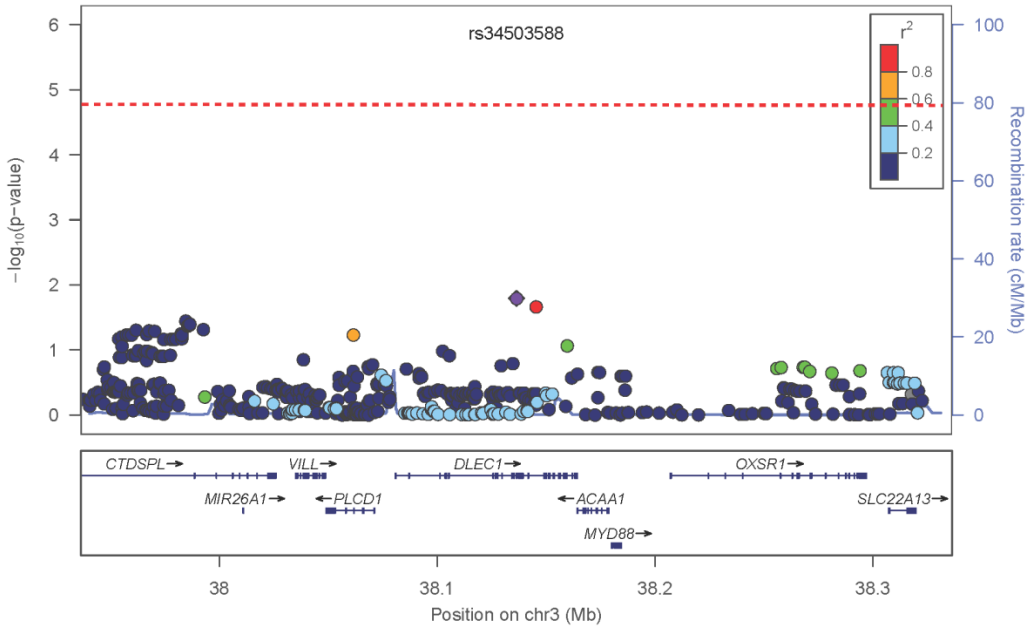


Figure 2-2. Manhattan plot showing association of analyses of SNPs in and near PLCD1

2.4 Discussion

2.4.1 *ATXN2*

I found an association of a 47-SNP haplotype spanning *ATXN2* and *SH2B3*, as well as replicating the original 15-SNP haplotype as reported by Lahut and Omur et al. (2013)[213], in multiple datasets confirming a significant *ATXN2* haplotype association with ALS in UK, Turkish and possibly Netherlands populations[213].

The original finding by Elden et al. (2010)[97] has been replicated in a wide range of populations since [204, 205, 208-210, 212, 222-224], confirming ATXN2 association with ALS. The association of an *ATXN2* haplotype in this thesis substantiates the original finding, as it confirms association in multiple ALS populations and with *ATXN2* polyQ repeat carriers directly[213].

It is possible that a specific *ATXN2* haplotype gives rise to polyQ expansions by causing regional instability. The fact that repeat lengths seem to be highly variable may be indicative of this. This

is a similar argument put forward to how the *C9ORF72* hexanucleotide mutation haplotype gives rise to the hexanucleotide mutation itself[225-227]. Although establishing a definitive *ATXN2* haplotype has been difficult, even across similar populations (for example USA and UK). It may be the case that (a) there is multiple founder effects for *ATXN2* repeat instability, (b) that there have been frequent recombination events that have caused divergence in *ATXN2* haplotypes, (c) that there is common variation of an *ATXN2* haplotype making to identify using GWAS methods, or (d) that the penetrance of *ATXN2*-caused ALS is too rare be identified using current GWAS methods and available ALS population sizes.

Specificity of this haplotype to the UK may be a consequence of a great marker density in the genotyping platform than other countries compared here. However a highly iterative imputation of genotypes was conducted using a method found to correlate very highly with the original raw genotypes [228].

An alternative explanation is that there is significant underrepresentation of the *ATXN2* polyQ repeat carriers, and therefore the associated haplotype, in non-UK populations. This is unlikely as the carriers have been shown to be proportional between USA, UK and Netherlands populations. It is more probable therefore that if there is a haplotype linked to the *ATXN2* repeat mutation then there is a clear founder effect specific to UK and Netherlands populations, but this scenario is unlikely due to genetic similarity between USA and UK Caucasian populations.

2.4.2 *PLCD1*

PLCD1 was found to have significantly associated with two populations (UK and Belgium) but not the USA[214, 215]. There is also statistically significant evidence of differential gene expression in *SOD1* mice[215]. Unfortunately I was unable to replicate the association using the dataset above, nor did I identify significant differential expression of PLCD1 in Chapters 5 and 6.

As Staats et al. (2013)[215] discusses, fine-mapping the microsatellite association in Belgium and UK populations show different associated alleles[214]. One possibility is that rare *PLCD1* point mutations are heterogeneous between populations, similar to the population heterogeneity of point mutations in *SOD1*. If this is true then the method of imputation and association analysis of genotypes is unlikely to uncover these mutations.

PLCD1 lies in a very gene dense neighbourhood. It is 22 kilobases (kb) long and has 15 coding exons, but unfortunately we had coverage of only one actual (non-imputed) genotype within *PLCD1*. Correlation comparing imputed and actual genotypes in this region was $r^2 = 0.87$, suggesting a reasonable accuracy of imputed genotypes. However this is an average and imputed genotypes in and around *PLCD1* may have been uninformative.

A third possibility for the null finding is that common variants may not be suitable for elucidating *PLCD1* pathology. For example, in *ELP3* there is linkage disequilibrium between the microsatellite marker and allelic markers outside of the *ELP3* gene. This may be capturing an ALS-linked *ELP3* haplotype but may also, for example, be elucidating an eQTL or repeat yet unexamined. As has been witnessed recently with *C9ORF72*, association with a genomic region in ALS does not immediately elucidate its genetic involvement in the disease.

If the lack of genetic association of PLCD1 with ALS is true, then PLCD1 involvement is likely due to post-translational processes. As it stands, it is therefore unlikely that PLCD1 is a causative gene of ALS.

2.4.3 Summary

Linkage, GWAS and sequencing analyses allow genetics to be examined through different magnitudes, progressing from large genomic regions to the signal nucleotide. But also, each method also carries along with them assumptions about the disease. Under the assumption that ALS disease regions co-segregate with the disease, linkage analyses are useful for familial

inherited diseases, which putatively make up 5% of ALS patients. There has been success in identifying genetic mutations using ALS families, although the genetic inheritance of ALS in these families are examined using somewhat Mendelian principles. Unfortunately the oligogenic and complex genetic contribution in ALS weakens the justification that we can take these Mendelian inherited large-effect size mutations and draw conclusions about the much larger 95% sporadic ALS population.

GWAS brings assumptions in line with the common variant common disease hypothesis, where an orchestra of common small effect-size genetic variants cause the disease. GWAS in ALS has been successful but less informative about genes outside of *C9ORF72*. This may be a power issue with ALS cases difficult to obtain and high heterogeneity within the disease. Many genes we know to be involved in ALS, for example *SOD1*, do not contain variants that are significant in ALS GWAS studies, despite having over 120 different point mutations, many of which are highly penetrance but rare. We should therefore not expect strong genotypic associations with ALS for genes that we believe have an involvement from other analyses.

I have examined two genes, *ATXN2* and *PLCD1*, using the same method of imputation and association testing of genotypes individually and as haplotypes. For *ATXN2* I elucidated a previously identified haplotype that associates with ALS. But the same method was unsuccessful for *PLCD1*. I believe this highlights the difficulty in identifying various types of genetic mutations involved in ALS. *ATXN2* harbours a repeat mutation and may support the involvement of a haplotype background, identifiable using genotype association methods. The data for *PLCD1* however currently suggests a more 3-dimensional modulatory alteration to genes functionality, unidentifiable using these methods.

Chapter 3 Residual association at *C9ORF72*

3.1 Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease of motor neurons that causes relentless paralysis, with death occurring within two to five years. The lifetime prevalence of ALS is 1 in 300 [229] with peak onset about 60, and an increased incidence in men [230]. Genetic studies of ALS have begun to yield insights, with ten moderate to high penetrance Mendelian risk genes now identified for adult onset ALS, with or without frontotemporal dementia (FTD): *SOD1* [231], *TARDBP* [44, 45, 232], *FUS* [59, 60], *ANG* [233], *OPTN* [73], *VCP* [76], *C9ORF72*[82, 83], *FIG4* [234], *UBQLN2* [3] and *PFN1* [235]. Other variations acting to increase risk may include intermediate expansion of the *ATXN2* gene [97], *ELP3* intron 10 variation , an *UNC13A* SNP [79], and indels in the *NEFH* tail domain [90, 91].

In about 5% of cases, a family history of ALS is recorded, with the remaining cases labelled sporadic [236], but this distinction is artificial, with an increased risk of other neurodegenerative diseases including Parkinson's disease, and frontotemporal dementia [72, 236] as well as an increased risk of ALS in relatives of those with apparently sporadic disease [237]. Furthermore, variants in every Mendelian ALS gene have also been reported in those with apparently sporadic ALS [4, 238], and the heritability of apparently sporadic ALS has been estimated at 61% [13].

Genome-wide association studies in ALS and FTD recently identified a locus on chromosome 9 within a known linkage peak for ALS-FTD [79-81]. The disease-causing mutation has now been identified as a massive expansion mutation of a hexanucleotide repeat between non-coding exons 1A and 1B of *C9ORF72* [82, 83]. Unaffected individuals have 2 to 23 repeats of the $(GGGGCC)_n$ hexanucleotide microsatellite, while the pathological repeat mutation may be several

hundred repeats. The mutation is interesting in manifesting with at least four phenotypes: ALS, ALS-FTD, FTD and normality.

Although three different genome-wide association studies all identified rs3849942 as the most associated SNP at the *C9ORF72* locus, and it is clear that this is in strong linkage disequilibrium with the repeat mutation [225], the UK sub-population showed stronger genome-wide significant association at a different SNP, rs903603. The most parsimonious explanation is that rs903603 and rs3849942 are in strong linkage disequilibrium and both are tagging the pathological expansion. If that is the case, one would expect that removal of cases with the repeat mutation from the analysis would obliterate the association at both SNPs. We therefore tested this hypothesis in the UK population.

3.2 Method

3.2.1 Sample collection

Whole blood samples from the MND DNA Bank were used. Selection was based on no family history of ALS, being of white European ancestry, and disease onset January 2002 or after. Control samples were obtained from the Depression Case Control (DeCC) study, the Bipolar Affective Case Control Study (BACCS), from Panos Deloukas of the Wellcome Trust Sanger Institute (Cambridge, UK) and the British 1958 birth cohort DNA collection. This project was ethically approved.

3.2.2 Sample preparation

DNA was extracted from blood samples by standard methods within one week of bleeding and stored at the UK DNA Banking Network in Manchester. Samples were bar-coded in a tracking system to reduce risk of clerical error.

3.2.3 Genotyping

Genotyping and quality control was carried out as described previously [80], but in brief, genome-wide association was performed by genotyping on various Illumina DNA microarrays followed by standard quality control and association testing using PLINK [220] .

The mutation was identified using repeat-primed PCR as described [82]. Non-mutation repeat length for alleles of the hexanucleotide repeat was quantified using Amplified Fragment Length Polymorphism (AFLP), using fluorescently-labelled primers as detailed previously. [82] Both analyses were performed using an automated 3130XL capillary electrophoresis-based Genetic Analyser and GeneMapper v4.0 software (Applied Biosystems).

3.2.4 Statistical quality control

Statistical quality control (see supplementary data) excluded individuals with incongruent sex, those marked for exclusion by the genotyping institution, those with low genotyping per individual <0.02, those with phenotype missingness <0.02, those with low heterozygosity ($p<0.05$), those exhibiting identity by descent sharing >5% of alleles with another participant, and those not of white European ancestry.

SNP markers were excluded with a minor allele frequency (MAF) below 0.015, SNP missingness <0.02, significant departure from Hardy-Weinberg equilibrium ($p < 0.001$), those marked for exclusion by the genotyping institution, and those on chromosome X.

3.2.5 Phasing and imputation analyses

Phasing and imputation was completed using 1000 Genomes v2.20101123 autosomal release (http://www.1000genomes.org/), performed using MaCH 1.0.18 (http://www.sph.umich.edu/csg/abecasis/MACH/index.html) and minimac (http://genome.sph.umich.edu/wiki/Minimac). We performed association analysis on imputed data using ProbABEL (http://www.genabel.org/packages/ProbABEL) , mach2dat (http://genome.sph.umich.edu/wiki/Mach2dat:_Association_with_MACH_output), and PLINK v1.07(http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml), after creating input files using the GenGen (http://www.openbioinformatics.org/gengen/) conversion tool. For GenGen, we set the r^2 threshold as 0.3 and, based on genotype posterior probabilities derived in Mach, a quality control threshold of 0.9.

3.2.6 Association and haplotype analysis

To test for residual association and identify SNPs not related to the mutation we used logistic regression, modelling case-control comparisons with sex as a covariate, stratified by the presence or absence of the repeat mutation.

We set the Bonferroni-corrected threshold of statistical significance at 1.13×10^{-7}, and the genome-wide threshold as 5×10^{-8}. We utilised the R statistical package to construct Manhattan

and Q-Q Plots, LocusZoom (http://csg.sph.umich.edu/locuszoom/) for regional Manhattan plots, and Haploview for haplotype block maps (http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview). To correct for genomic inflation, p-values were adjusted by the median χ^2 statistic of the genomic inflation factor.

For linear regression analysis of the non-mutated microsatellite repeat lengths as a quantitative variable we used PLINK v1.07 and IBM SPSS 20. For haplotype analyses we used PLINK v1.07 sliding windows between 2 and 100 to explore significantly associated haplotypes using the phased imputed dataset. The prevalence of the most associated haplotypes were refined in Haploview and tested for their association with ALS using an omnibus haplotype association analysis.

We ran all genetic tests under additive, dominant and recessive models, with little difference in results between dominant and additive models. We have reported the additive model for association of SNP alleles with the repeat mutation, and a dominant model (using the larger allele) for the analysis of repeat lengths.

An epistasis analyses utilised a logistic regression interaction coefficient, examining cases-only through a synthetically created fakeSNP (representing the HREM) against all SNPs genome wide. The fakeSNP (rsFAKE10592147) was integrated into the dataset at chromosome 9 at base position 27,563,540, the estimated start point of the HREM, close to rs10592147 (HG18, Genome Reference Consortium Human Build 36 (GRCh37)). The SNP was coded homozygous for carriers and heterozygous for non-carriers. To confirm that this did not violate assumptions of an additive

model we implemented the HREM status into the model as a bivariate covariate and obtained identical results.

3.2.7 Sequencing *C9ORF72*

Four hundred and eighty DNA samples of patients with sporadic ALS were pooled in sets of eight. Each pool was tagged with one of 12 identifying DNA sequences and then combined into a further pool containing one representative of each identifying sequence. Thus each final sample for sequencing contained the DNA of 96 different individuals which could be resolved down to a group of eight. Samples were sequenced using Illumina GAIIx platform generating 38 base-pair paired-end, indexed reads. Preliminary analysis and QC of reads was carried out using Illumina's CASAVA v1.7 software and output files contained all filtered, but unaligned, sequences for each sample stored in a simple exportable FASTQ format. Alignment of short reads against reference mRNA sequence (uc003zqq.2, *C9ORF72*, length=3233) and variations calling was done using maq 0.7.1 mapping short DNA sequencing reads and calling variants using mapping quality scores with default parameters [239].

3.2.8 Sequencing the repeat

Forward (GGTTTAGGAGGTGTGTGTTTTTGT) and reverse primers (CCAGCTTCGGTCAGAGAAAT) were designed to create a 424bp amplicon (64% GC-rich) covering the repeat sequence. A touchdown PCR protocol (see protocol P1 in appendix) was used to amplify the region of interest followed by gel electrophoresis to inspect the amplicon length.

Samples were purified and analysed using an ABI 3130xl Genetic Analyser (Applied Biosystems) with the BigDye Terminator v1.1 Cycle Sequencing Kit. ABI's Sequencing Analysis 5.3.1 software was used to analyse base calls and identify repeat patterns.

3.3 Results

3.3.1 *C9ORF72* genotyping

There were 632 cases and 4,519 controls. After quality control there were 599 cases and 4,136 controls. There were 585,919 SNPs; after quality control 442,057. 39 cases tested positive for the mutation; the remaining samples had repeat lengths of less than 23 (Table A2 and data analysis D1 in appendix).

3.3.2 Residual association

Genomic inflation λ_{GC} was 1.00 – 1.03 for all analyses. Genome-wide association of all UK cases vs. controls confirmed the findings of the original study[80], with no new associations (Figure A2 and Table A3). The top three SNPs were rs10967976 (p-value = 6.164 x 10^{-8}; OR: 1.41), rs903603 (7.548 x 10^{-8}; OR: 0.71), and rs10812611 (9.819 x 10^{-8}; OR: 1.4), all at 9p21.2. rs903603 was the most associated SNP in the previous independent analysis [80]. In that study rs10967976 and rs10812611 had been removed for compatibility with the other platforms used in the joint analysis. To confirm that the association signal seen was a result of the mutation and that there were no new associations, we performed association testing excluding the 39 mutation positive cases (Figures A3 and A4, and Table A4). There was a residual association at 9p21 with the same top three SNPs; rs903603 (1.87e-05, OR: 0.76), rs10967976 (3.63e-05; OR: 1.31) and rs10812611 (5.02e-05; OR: 1.30), suggesting that despite being at the disease locus, they are not tagging the pathological repeat mutation. This is reflected in the change in p-value ranks with the most

associated SNP from our original study, rs903603, remaining one of the ten most associated SNPs after removal of cases with the mutation, compared with the SNP tagging the repeat mutation, rs3849942, dropping from 7th to 1,225th place, as expected (Table 3-1).

	With hexanucleotide repeat mutation			**Without hexanucleotide repeat mutation**		
SNP	**Rank**	**P-Value**	**Odds ratio (allele)**	**Rank**	**P-Value**	**Odds ratio (allele)**
rs10967976	1st	6.16E-08	1.41 (G)	23rd	3.63E-05	1.31 (G)
rs903603	2nd	7.55E-08	0.71 (A)	8th	1.87E-05	0.76 (A)
rs10812611	3rd	9.82E-08	1.4 (G)	34th	5.02E-05	1.3 (G)
rs3849942	7th	3.28E-06	1.39 (A)	1225th	2.57E-03	1.25 (A)
rs2814707	9th	5.00E-06	1.38 (A)	1562nd	3.30E-03	1.24 (A)

Table 3-1. Change in Rank by P-value for Residual and Mutation-Specific SNPs

We confirmed that the mutation tagging SNP rs3849942 is strongly associated with ALS when only samples positive for the mutation were tested against controls ($p = 9.77 \times 10^{-12}$, OR: 5.225, rank = 1), while rs903603 was not as strongly associated as rs3849942 ($p = 7.73 \times 10^{-7}$. OR 0.24, rank =22) (Figure A5). A QQ plot (Figure A6) did not show evidence of inflation of the test statistic.

3.3.3 Epistasis analysis

The genome-wide epistasis analyses revealed several candidate regions showing statistical interaction with the repeat mutation. Note that these are not corrected for multiple testing and that the bonferroni line is 1.12×10^{-7}. None of the associated SNPs were the residual SNPs found in the previous section. The synthetic SNP in our dataset (rsFAKE10592147) represented HREM presence in 39 cases and the 560 non-HREM cases. Controls were excluded. Examining statistical interactions between rsFAKE10592147 against all SNPs genome wide resulted in three of these SNPs being the top hits (Table 3-2). We noted four none chromosome 9 SNPs from comparing HREM only with controls, and HREM-cases with non-HREM cases, that nearly made genome-wide significance in the previous analyses. These were found the top SNPs (highlighted in Table 3-2). Note that the residual association logistic regression parameter and the interaction parameter taken from this epistasis regression model are very similar. This result highlights possible trans-acting epistatic interaction.

Chromosome A	Fake SNP	Chromosome B	SNP	Statistic	P-Value
9	rsFAKE10592147	17	rs218679	24.41	7.81E-07
9	rsFAKE10592147	10	rs2499081	20.05	7.54E-06
9	rsFAKE10592147	22	rs12165820	19.93	8.03E-06
9	rsFAKE10592147	9	rs3763662	19.19	1.18E-05
9	rsFAKE10592147	17	rs9910747	18.72	1.51E-05
9	rsFAKE10592147	8	rs2294011	18.63	1.59E-05
9	rsFAKE10592147	16	rs886434	18	2.21E-05
9	rsFAKE10592147	10	rs11198878	17.98	2.23E-05

Table 3-2. Genome-wide epistasis analysis with rsFAKE10592147 representing the HREM

3.3.4 Residual association and mutation-specific haplotypes

To explore linkage disequilibrium in the region, we phased and imputed SNPs at *C9ORF72* and the surrounding region (chr9:27303051-27750375) using the 1000 Genomes Project reference panel. Haplotype association analyses and refinement of haplotype blocks in Haploview identified a 79-SNP haplotype in 100% of the cases with the mutation (n =39), of which 72 SNPs were from

an 82-SNP haplotype published previously [225]. Using the same method we also identified a distinct 11-SNP haplotype present in 76% of the samples without the mutation (Figure A7).

3.3.5 Repeat length and risk allelic distribution

We examined the relationship between the hexanucleotide repeat length of phased cases without the mutation (n=448) and the allelic frequency and distribution of ALS-associated SNPs to explore the linkage disequilibrium pattern of the locus. The mutation-tagging SNP rs3849942 risk allele was associated with increased hexanucleotide repeat numbers for the larger allele in linear regression (β = 5.711, p = 3.02×10^{-84}) (Table A5). Allele frequency distributions of the repeat length stratified by rs3849942 alleles showed a marked difference between those with six or less against seven or more (Figure 3-1). Non-mutation samples with seven or more repeats (n=208) showed the strongest genome-wide association with ALS when compared to controls (n=4142) OR: 4.99; p = 1.35×10^{-43}) (Table A6).

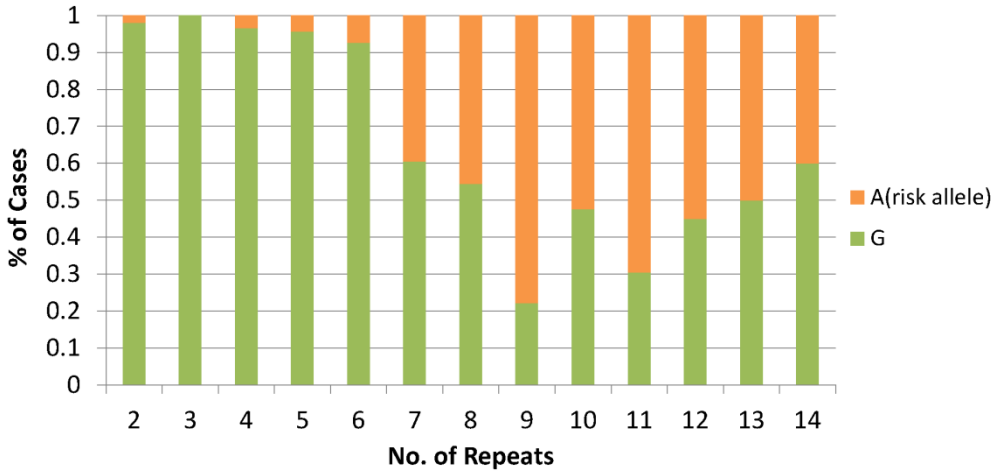


Figure 3-1. Relationship between rs3849942 alleles and repeat length in non-mutation cases

The risk allele is strongly associated with longer repeat lengths, suggesting it may lie on a haplotype promoting repeat length instability.

A similar relationship between risk alleles and hexanucleotide repeat length was observed for the SNPs showing residual association. Repeat numbers greater than two (the minimum length) for the larger allele were significantly associated with risk alleles at all of the three SNPs showing residual association (Figure 3-2). Modelling repeat length as a quantitative variable in a linear regression confirmed that these SNPs were significant predictors of having greater than two repeats (example SNP rs10812611, $\beta = 5.835$, $p = 1.50 \times 10^{-52}$) (Table A5). Again, association testing of samples with repeat length greater than two (n=385) against all controls (n=4142) showed that SNPs showing residual association were most significantly associated with ALS; rs903603: OR: 0.39; $p = 2.21 \times 10^{-28}$, rs10812611: OR: 2.78; $p = 6.96 \times 10^{-33}$, and rs10967976: OR: 2.82; $p = 1.19 \times 10^{-33}$ (Table A7).

The identified 11-SNP haplotype for the residual association SNPs was found in 100% of the samples with greater than two repeats for larger allele (71% in controls). Extension of the haplotype to include a further SNP, rs2492816, was very specific for longer repeat lengths. This 12-SNP haplotype was strongly associated with ALS in the sample set excluding cases with the mutation, or with two or less repeats ($\chi^2 = 114$, p-value = 1.56×10^{-24}, Figure 3-3). This haplotype was not significantly associated with ALS in cases with two hexanucleotide repeats and was not found at all in cases with the pathological repeat mutation.

An omnibus association analysis comparing mutation-specific and residual association haplotypes in those with the mutation and those with greater than two repeats on their larger allele supported the hypothesis that the alleles showing residual association at *C9ORF72* are significantly associated with ALS, but not by association with the pathological repeat mutation (Table 3-3).

	Haplotype p-value for association with ALS	
Haplotype:	**Mutation-specific**	**Residual**
Mutation Cases (n=39)	**7.53×10^{-14}**	1.30×10^{-8}
Cases with hexanucleotide repeat length >2 (n=385)	9.87×10^{-9}	**5.91×10^{-23}**

Table 3-3. Association analyses of haplotypes with ALS in mutation cases and cases >2 repeat

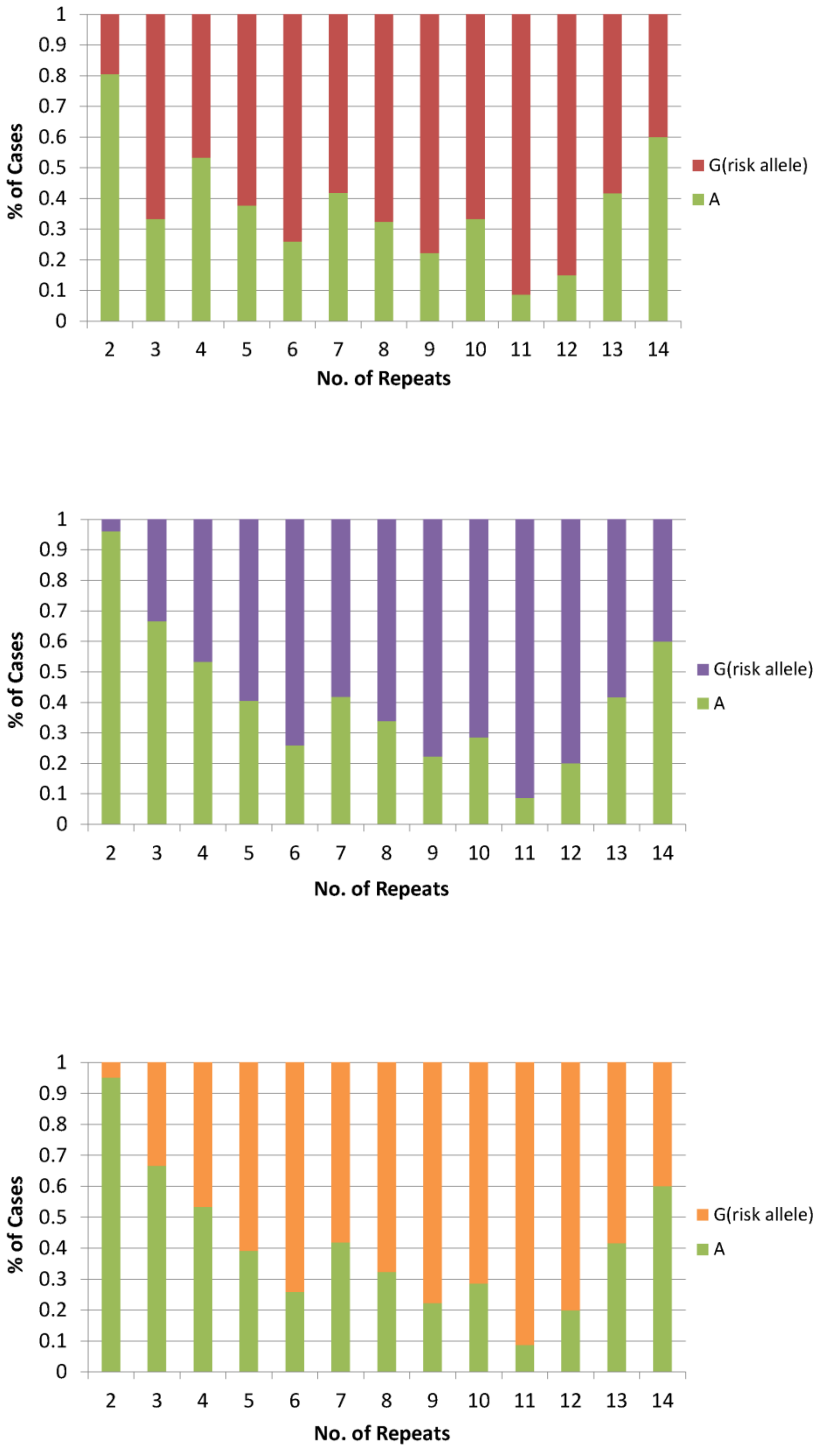


Figure 3-2. The relationship between hexanucleotide allele repeat length and SNPs showing residual association at the *C9ORF72* locus.

SNP rs903603 (top), rs10967976 (middle), and rs10812611 (bottom). Just as for the relationship for rs3849942 shown in Figure 3-1, the risk allele for SNPs on the alternative risk haplotype is overwhelmingly likely to be associated with repeat sizes greater than two.

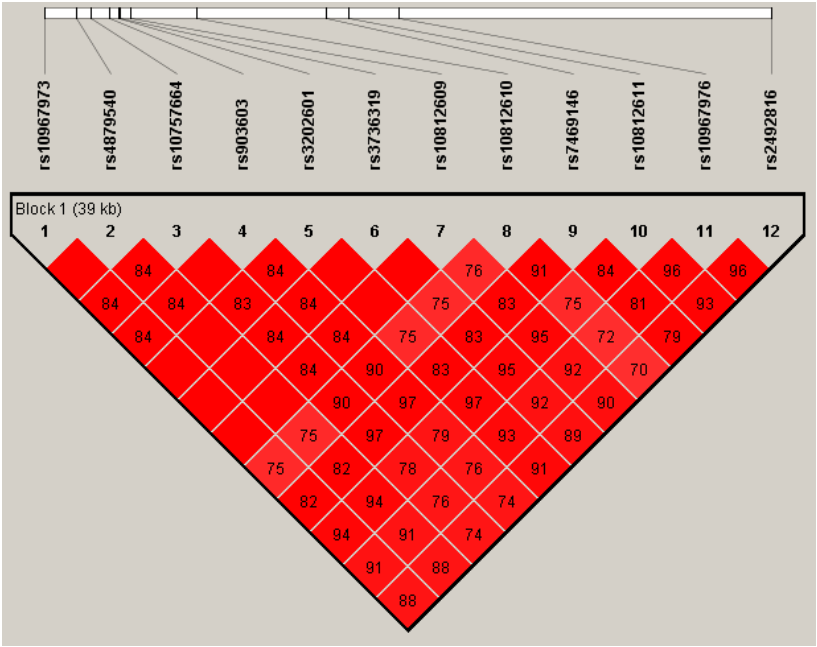


Figure 3-3. 12-SNP haplotype for cases with repeat length greater than size two

3.3.6 Apparent homozygosity for larger hexanucleotide repeats associates with the risk allele of rs903603

The association of risk alleles with greater than two hexanucleotide repeats suggests that residual association SNPs might predispose to instability, and therefore a repeat mutation, as has been postulated previously for the rs3849942 haplotypic background. Thus, one possible explanation of the residual association is an interrupted or alternative repeat sequence, also pathologically expanded, but not detected by repeat-primed PCR targeted at $(GGGGCC)_n$ (Figure 3-4). In that situation one would expect a similar problem to that reported by DeJesus-Hernandez: that cases appear homozygous because their larger repeat allele is too expanded to amplify. 93.8% (15/16) of cases with greater than two repeats and the rs903603 risk allele were apparently homozygous

for the hexanucleotide repeat allele, breaking Hardy-Weinberg equilibrium. In comparison, 66% (150/266) of cases with the same risk allele were heterozygous for the repeat and we found only 4% (3/83) of cases with the risk allele homozygous for 2 repeats (Table 3-4). We did not observe this >2 repeat homozygosity bias for any other SNPs. For example, 37% of cases with the mutation-specific risk alleles had homozygosity for repeat alleles of length greater than 2, which satisfied Hardy-Weinberg equilibrium.

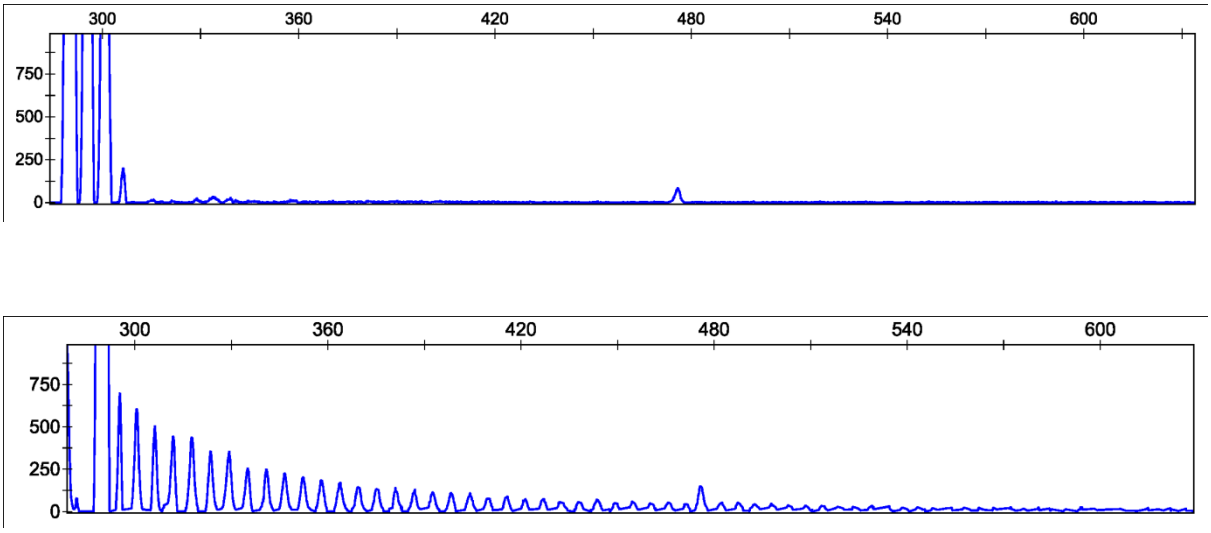


Figure 3-4. Repeat-primed PCR results.

Repeat-primed PCR of example cases showing a small expansion in the non-pathological size range (top), and a sample known to have a pathological $(GGGGCC)_n$ expansion mutation (bottom).

Repeat Zygosity	**Non-risk Allele A (%)**	**Risk Allele G (%)**
Heterozygous	76 (34%)	150 (66%)
Homozygous 2 repeats	80 (96%)	3 (4%)
Homozygous (repeats >2)	1 (5%)	18 (95%)

Table 3-4. Frequency of cases by residual rs10967976 alleles stratified by repeat length and zygosity

3.3.7 Trimodal pattern of repeats

We found a trimodal pattern of repeat allele sizes, where the occurrence of 2, 5 or 8 repeats was significantly more frequent than other repeat numbers (Figures A8 and A9). These findings suggest some repeat sizes are more stable than others.

3.3.8 Sequencing Analysis of *C9ORF72*

We identified two point mutations, a ketone G/T and an amine A/C substitution, across 31 (n= 272) and 2 (n= 16) pools, respectively. No group analysed together or separately showed association with the residual or mutation-specific risk haplotypes. There was also no sample showing homozygosity for repeat sizes larger than 2. Thus it is unlikely that the residual association identified is a result of the point mutations seen.

3.3.9 Sequencing Analysis of the *C9ORF72* repeat

We sequenced the repeat region in 56 samples: 18 cases homozygous for greater than two repeats with the risk haplotype showing residual association, nine homozygous for repeats of size two, 10 heterozygous for repeat length, and a mixture of non-ALS controls homozygous for repeat sizes greater than two (n = 6), homozygous for repeats of size two (n = 7), and heterozygous for

repeat length (n = 6). Our previous length and zygosity estimates based on AFLP (Figure A10) and repeat-primed PCR were correct for all these samples.

3.4 Discussion

We have shown that after removing cases with expansion mutation of the (GGGGCC)n repeat at *C9ORF72* there remains residual association of SNPs at this locus with ALS. As expected, the SNPs associated with the mutation are no longer associated with ALS in this sample, but rs903603 and SNPs in strong linkage disequilibrium with it are associated, both before and after the removal of cases with the mutation. These SNPs showing residual association reside on a distinct haplotype, different from that of the mutation, and like the mutation-tagging SNPs, are also associated with non-pathological expansion of the hexanucleotide repeat, suggesting that they too promote instability of a repeat sequence.

Because the SNPs showing residual association are not associated with the mutation but are associated with ALS, repeat instability, and apparent homozygosity for the repeat, one possible explanation is an alternative pathological expansion that is not detected by repeat-primed PCR targeting (GGGGCC)n. Such an alternative repeat sequence might be the result of an interrupted repeat. Interrupted repeat sequences are found in ALS resulting from intermediate expansion of a CAG trinucleotide repeat in the *ATXN2* gene, with CAA interrupting the sequence and resulting in an alternative [240]. Interrupted sequences are also found in Fragile X syndrome, a condition with a massive pathological repeat expansion [241].

Others have reported evidence of pathological *C9ORF72* hexanucleotide expansions by Southern blotting without corresponding detection of repeat-primed PCR in FTD cases, consistent with an alternative repeat sequence [242].

Current evidence supports the idea that the sheer size of the disease-associated *C9ORF72* hexanucleotide repeat expansion causes pathology, with no evidence for repeat lengths below

23 being associated with ALS. One would expect that an alternative repeat that is also pathological should also be very large and not readily amplified by PCR. In that case, samples carrying the alternative repeat should appear homozygous for the non-expanded allele since the pathological allele would appear null by PCR. Consistent with this, we find cases with more than two repeats are overwhelmingly likely to appear homozygous.

One factor increasing the likelihood of pathological expansion in transmission to the next generation is the current size of the repeat, and it is particularly interesting that the SNPs showing residual association also strongly associated with alleles showing larger numbers of non-pathological repeats. The *C9ORF72* region is very GC-rich, known to increase vulnerability towards repeat mutations. Additionally, a familial study examining inheritance of the repeat mutation suggests that larger repeat lengths are indeed more unstable and prone to expansion [227].

We found a trimodal pattern of repeat allele sizes with significant departure from a uniform distribution for repeat lengths 2, 5 and 8. This pattern is also apparent in previous publications [82, 225], although not explicitly described. The tendency to increase the allele size by 18 nucleotides might be related to the underlying pathological mechanism of expansion, either because the mechanism results in preferential expansion in units of 18 base pairs, or because expansion in other unit sizes is unstable and leads to massive pathological expansion.

An epistasis analyses helped confirmed the presence of several non-9p21 SNPs that statistically associated with the HREM cases compared to non-HREM cases and controls. The most significant was rs218679 which is found in intron 1 of solute carrier family 13, member 5 (SLC13A5), involved in the sodium-dependent transport of citrate in cellular entry and functionally involved in circulating citrate for metabolic energy [243]. The second epistatic SNP was rs2499081 found in the intergenic region before c10orf112, which carries the Alzheimer's risk marker D10S1423 that is likely to affect the gene through LD [244]. The third epistatic SNP was rs12165820 in intron 1 of *FAM19A5*, which (intron and gene) has been implicated in ALS previously in Schymick et al. [245] GWAS and in a GWAS study presented as an abstract at the Annual Meeting San Francisco 2008 (http://www.ashg.org/2008meeting/abstracts/fulltext/f22492.htm), both through the identification of rs130110, 129kb away from rs12165820. *FAM19A5* is a member of

the TAFA family and its putative function is in brain specific neurokines that help regulate immune and nervous cells. Rs130110 was not significant in our analyses and LD between rs130110 and rs12165820 was $r^2 = 0.006$ and $D' = 0.110$.

A weakness of this study is that we relied on repeat-primed PCR to identify cases with the mutation. It is possible that some cases with the mutation were missed and this explains the residual association. Several factors make this unlikely. First, the mutation is tagged by rs3849942, and as expected, this SNP lost association with ALS after the removal of mutation-positive cases, while rs903603 remained in the top ten associated SNPs. Second, the frequency of identified mutation cases in our sample is among the highest in the world [225], suggesting we are not missing a large proportion of these. Third, we performed a standard PCR of the hexanucleotide repeat and assayed the length polymorphism of the alleles (AFLP); in all cases, there were no heterozygous cases with the expansion, suggesting that the detection method worked. Fourth, in all cases where we had sequence data, the repeat size on sequencing matched our estimate of size based on AFLP and repeat-primed PCR.

Another explanation for our findings is a substitution or similar mutation, rather than an alternative repeat mutation. This seems unlikely because for several years many groups, including ours, performed sequencing through the region finding no segregating point mutations or indels. Furthermore, the point mutations we identified on sequencing do not occur in individuals carrying the alternative risk haplotype. It does however remain possible that for a small subset of individuals with sporadic ALS, point mutations of *C9ORF72* are pathogenic, but it is then difficult to explain the association between rs903603 risk alleles and hexanucleotide repeats greater than length two. Nevertheless, we cannot exclude this possibility.

A third possibility is that SNPs showing residual association tag repeat alleles greater than size two because such repeats are themselves pathogenic. Cases and controls share the same frequency and distribution of repeat lengths [246] making this unlikely, but recent evidence suggests that moderate expansion in the number of repeats can actually be pathogenic [247].

A fourth weakness is that I was unable to replicate the residual association in non-UK populations (Netherlands, Ireland, and USA). Statistical power was not an issue (except for Ireland) and we

do find the statistical association of rs3849942. It is possible that the residual association exists in the UK only, due to second founder event. Alternatively the haplotypes between countries may be different and was not captured combining the genotypes from their various platforms.

We present evidence of residual association at the *C9ORF72* locus in ALS, not accounted for by the known pathological GGGGCC hexanucleotide repeat expansion but consistent with the hypothesis of an interrupted GGGGCC or alternative sequence repeat. We present a new pathological haplotype associated with ALS at this locus that is associated with having greater than two repeats for the larger allele, suggesting this haplotype predisposes to repeat instability. Further investigation including Southern blot analysis of individuals carrying the risk haplotype and sequence analysis of the region will help to identify the cause of the residual association signal we have identified.

Chapter 4 Genetic analyses using RNA-sequencing: *ELP3* and Chromosome 9p21

4.1 Introduction

The genetic work in this and the following chapter builds upon previous findings from collaborators or publications. The genetic regions examined cover the Elongator Protein 3 (*ELP3*) gene and the intergenic region between chromosome 9 open reading frame 72 (*C9ORF72*) and Mps One Binder kinase activator-like 2B (*MOBKL2B*); which are both detailed in the literature review.

The analysis below uses pre-existing RNA sequence (RNA-seq) data for generating hypotheses on the pathological causes of the ALS, and therefore upcoming research. The RNA-seq data comprises of control and Alzheimer brain tissue. Mainly bioinformatic approaches were used to examine any unusual variation in these genes, mainly by examining alternative transcripts, which may give us clues to their contribution to ALS. I then followed these up using wet-lab methods.

Unfortunately both projects ended before completion; *ELP3* due to grant completion and *C9ORF72* due to the hexanucleotide expansion mutation discovery[82, 83].

4.1.1 *ELP3*

ELP3 was implicated in ALS through a UK GWAS using microsatellites[214]. From 1884 microsatellite markers, four markers were selected based on their association statistical significance and candidacy. Using Monte Carlo based permutation for the multiple alleles of each marker, STS microsatellite marker D8S1820 significantly associated with ALS in UK and Belgium populations, but not USA. D8S1820 is found in intron 10 of *ELP3*. 10 alleles in this marker, when grouped, have an OR of 2.07 of causing risk of ALS.

The authors searched for a functional variant in 61 possible SNPs, showing linkage disequilibrium with D8S1820 in three international populations. No convincing variant was found, nor in association studies following the original study, with the exception of Kwee et al. (2012) in US Veterans[248].

The original paper[214] also reported differential expression of the *ELP3* protein between ALS cases and controls, as well as stunted motor axonal growth after *ELP3* knock-down in Zebrafish. Overall these findings suggest a definitive role in motor neural growth and an involvement in ALS pathology.

As there is a genetic association of *ELP3* with ALS, but no analysis has provided convincing findings in terms of DNA mutation or abnormal gene expression, except for indirect evidence of lowered expression of the *ELP3* protein in ALS patients and its knockdown has aberrant effects on motor axonal growth, I adopted an approach looking at transcription variation. My method was to use RNA-seq data from non-ALS samples to identify alternative isoforms which may act as possible candidates for *ELP3* pathology.

4.1.2 Chromosome 9p21

Before the HREM was discovered in *C9ORF72*[82, 83], we were aware only, through linkage analysis of "9p21-families" and large-scale international GWAS[79-81], that a locus at chromosome 9p21 associated with a significant proportion of ALS cases. Several research groups had sequenced the DNA and RNA of *C9ORF72* with a high coverage depth (unpublished). *C9ORF72* was the most likely candidate in this region, but at the time neighbouring gene *MOBKL2B* and Interferon, Kappa (*IFNK*) were also candidates as the linkage area extended to cover this region. These genes too were sequenced extensively. Differential expression analyses of these genes showed little difference between cases and controls, despite selecting cases using the proposed 9p21-disease haplotype (which was somewhat similar to the final HREM haplotype) (unpublished).

I therefore proposed to look at intergenic regions using the RNA-seq data which may uncover single exon genes, aberrant non-coding RNA, cryptic exons, or a fusion site between *C9ORF72* and *MOBKL2B* (as the haplotype spanned across this region). Our main target was to examine rs3849942 and rs2814707, the top two and consistently associated SNPs with ALS from the international GWAS. These SNPs lie in the intergenic region between *C9ORF72* AND MOBK2LB. My approach was to search for erroneous expression in the RNA-seq data that could be linked to these SNPs as eQTLs. This project was disbanded in November 2011 following the *C9ORF72* hexanucleotide mutation discovery.

4.2 Method

4.2.1 RNA-seq samples and sequencing

Brain tissue from the MRC Brain Bank was used from four healthy individuals and five with Alzheimer's disease (AD), from two different brain regions- Entorhinal cortex (EC) and Brodmann area 9 (BA9). EC is affected early on and severely in AD; BA9 in the course of AD. The tissue was homogenised and RNA extracted by commercially available kits (Qiagen) and converted to cDNA by standard methods. The yield was approximately 100 micrograms of RNA from 100mg of tissue. Next-generation sequencing of samples was performed using an Illumina Genome analyser by Benjamin Blencowe from the University of Toronto. RNA-seq outputs were generated in the form pooled sequenced fragments of 36 base pairs long. I used the nine non-Alzheimer controls, to identify non-referenced variations around ALS candidate genes with the nine Alzheimer cases which I used as a cross-reference.

4.2.2 Sequencing alignment and quality control

The RNA-seq files were in Illumina fastq format, and were converted to Sanger fastq format using Perl script and FASTQC. This was primarily to convert headers for downstream tools, and to have quality score ranges into ASCII format.

Quality control was performed using FASTQC, which provides analyses of per base sequence quality, per sequence QC scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, duplicate sequences, overrepresented sequences and overrepresented k-mers.

To align reads I used TopHat[249] and Bowtie[250] to manipulate the reference genome HG18 (see Script A6 in Appendices). The output from TopHat is an "accepted_hits in a binary Sequence Alignment/Map" (BAM) format, which is a list of successful read alignments, and an exon-intron junction file and DNA insertion file, both in Browser Extensible Data (BED) format.

SAMTools[251] functions were used to manipulate these files; sorting and indexing were the two main functions used (see Script A6 in Appendices). IGV Tools were used to view the BAM files.

The MarkDuplicates function in Picard Tools (http://picard.sourceforge.net/) was used to identify duplicate reads.

At the time there was a compatibility issue between TopHat BAM files with GATK[252] and Picard Tools, so I used Stream Editor (SED) to alter header BAM header details (see Script A6 in Appendices).

4.2.3 Sequencing Analysis Methods

Two methods were used for variant calling; (a) using SAMtool mpileup -> BCFTools -> VCFTools[253], and SAM varFilter to identify high quality SNPs, and (b) GATK SNP calling function (see Script A6 in Appendices).

For transcript abundance estimates, regulation and differential expression analyses, using FPKM and RPKM, I used Cufflinks [254]. This tool produces Gene Transfer Format (GTF) files, in which can be used in Cuffcompare to compare transcript abundance in the data between genes or against a reference GTF file (used here from the UCSC). Cuffdiff was not used here as I was not comparing case and controls (see Script A6 in Appendices).

To identify novel splice junctions I used (a) SpliceMap[255] (see Script A6 in Appendices)and (b) TopHat without a reference genome

To identify and analyse expression of alternative transcripts not listed on the reference transcriptome I used MISO (see Script A6 in Appendices).

4.2.4 Genotyping

For genotyping of Sheffield University's Institute for Translation Neuroscience (SITraN) cases I used Invitrogen's TaqMan genotyping with an ABI 7900 HT Real-Time PCR system. I used TaqMan Genotyper to make allelic discriminations for cases and controls. Primers were designed and purchased using Eurofins MWG Operon primer design tool.

4.3 Results

4.3.1 *ELP3*

Initial results using Splicemap found previously unreported novel splice junctions; these can been seen in Figure 4-1 below. The Y-axis corresponds to the novel splice junctions found (in red), the X-axis is the *ELP3* gene (in blue) and below this is, are known transcripts (in black). We found approximately 20 transcripts with various abundance (not shown). The Ensemble database calculates *ELP3* to have 16 different transripts. There are (a) multiple known and possibly unknown *ELP3* transcripts and (b) our data shows novel splice junctions leading to transcripts not previously reported in the human genome reference dataset NCBI36/HG18. These transcripts are unlike the other *ELP3* transcripts, being shorter with one transcript ending not at a transcription codon stop sequence. This was found in seven out nine controls examined. MISO revealed that the abundance levels were low.

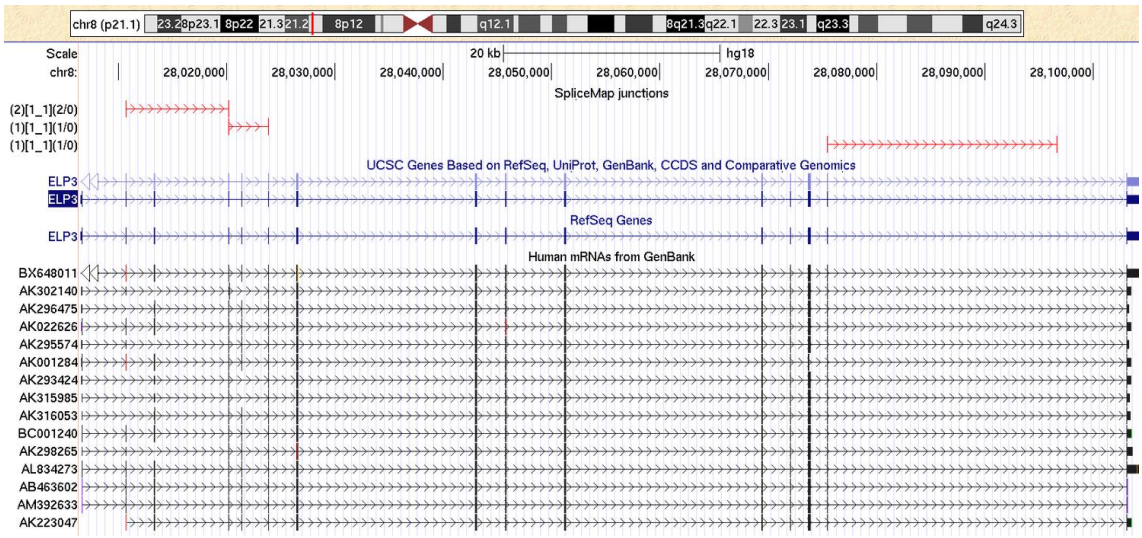


Figure 4-1. *ELP3* novel splice junctions detected using RNA-seq mapped to the UCSC Genome Browser

4.3.2 Chromosome 9p21

I analysed reference and de novo transcripts derived from Tophat using the RNA-seq data. I came across small sections of RNA expression in the intergenic region between MOBK2B and *C9ORF72*. On closer inspection I found that the intergenic transcripts lay close to rs3849942, the most significantly associated SNP with ALS worldwide. I examined this in relation to the HG18 reference genome using UCSC Genome Browser (see Figure 4-2), and found reported expression sequence tags (ESTs) in the same location. I gauged that this region might be a good candidate because either rs3849942 or rs2814707 could be acting as an expression trait locus (eQTL) in this region.

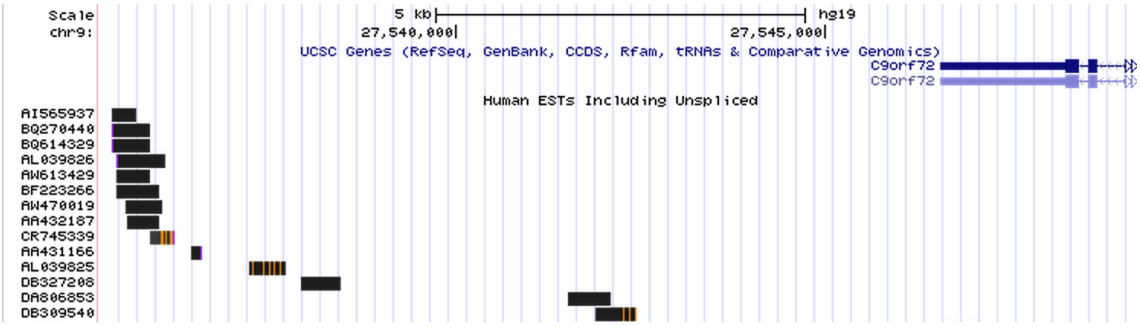


Figure 4-2. ESTs near rs3849942 near the 3' end of *C9ORF72*

I genotyped 124 samples from SITraN using ABI's TaqMan for SNPs rs3849942, rs2814707, and rs10122902. These three SNPs are known to form the 9p21 haplotype highly associated with ALS. Identifying ALS cases with this haplotype would isolate cases carrying the yet unknown mutation. During the course of this project the variant causing the association signal was identified as a hexanucleotide repeat expansion mutation in intron 1 of *C9ORF72*[82, 83], thus making the hypothesis I was testing unlikely to be true. Further testing was therefore abandoned.

4.4 Discussion

4.4.1 *ELP3*

Replication of the *ELP3* association with ALS has been supported by Kwee et al. (2012)[248], who found a SNP (rs6985069) located near *ELP3*, showing a significant albeit small association with ALS. Kwee et al. highlight the difficulty of accounting for genetic disease variance using common variants. It is becoming increasingly likely that they will not greatly inform the future of ALS research. Many large-scale multi-population genome wide analyses examining common variants have so far yielded only one convincing finding, *C9ORF72*, despite many attempts. The majority of findings in ALS so far have been rarer point mutations or repeat mutations.

Notwithstanding, *ELP3* does seem to have a role in ALS. Differential *ELP3* protein expression between *ELP3* risk-carriers and protective-carriers is evident. I hypothesised that *ELP3* pathology may be due to some transcriptional disruption without a definitive genetic mutation, for example an eQTL or aberrant transcription factor.

I explored alternative isoforms in healthy subjects to select candidate transcripts. I identified three rare alternative transcript isoforms in the majority of the samples, which are uncharacteristically like the other isoforms in the reference transcriptome for *ELP3*.

Interestingly there is one reoccurring isoform with moderate abundance incorporating many *ELP3* exons buts fails to transcribe to the final exon. This isoform stops short in the middle of the final intron and not at a termination codon sequence. The other two isoforms are short, but have their 5' and 3' ends in exons containing promoter sequences.

In comparison with other genes in our data and on the transcriptome, *ELP3* transcription is complex and is possibly vulnerable to aberrant splice formations. Future research should seek to characterise and measure alternative isoforms between cases and controls using qPCR as well genotyping rs6985069 and SNPs in LD to assess their eQTL status.

4.4.2 Chromosome 9p21

The haplotypic background of the *C9ORF72* hexanucleotide mutation, now believed to explain the original 9p21 linkage and GWAS signals, has been greatly detailed[225, 256]. rs3849942, the SNP examined in this thesis, is now known to be direct marker of the mutation, with 100% of mutation cases having the risk allele (A). rs3849942 and its associated haplotype may even give rise to the hexanucleotide mutation itself[225]. Nevertheless, with this findng the project was determined obsolete.

My findings suggested that the intergenic region between *C9ORF72* and *MOBKL2B* showed definitive areas of DNA sequence that were being transcribed, and which lay close to rs3849942 and the final exon of *C9ORF72*. More intriguingly was that the sequence showing transcription in our data was characterised by ESTs. ESTs are sequences of DNA believed to transcribe, usually, messenger RNA (mRNA). If rs3849942 was modulating intergenic mRNA in this region then this may interfering with the normal transcription of *C9ORF72* or *MOBKL2B*. I genotyped cases and controls in SITraN for rs3849942 and rs2814707; the project ended before I could move on to using qPCR to measure the largest EST in the region.

A benefit to analysing C9ORF72 expression using this method would be (a) to elucidate whether there is differential expression between cases and controls without the repeat hexanucleotide mutation and (b) to clarify whether one transcript variant is pathogenic or whether it is all.

It is possible that epigenetic events may reduce *C9ORF72* expression leading to a similar pathogenicity comparable to repeat-mutation carriers. Research examining this has so far been null {Belzil, 2013 #1456}. Furthermore, which transcript variants are affected by the mutation remains unclear. There are 3 variants, and there has been evidence that the mutation reduces transcript variants V1 and V2 but not V3, however *C9ORF72* protein levels were normal {DeJesus-Hernandez, 2011 #867}{Gijselinck, 2012 #931}. Other studies show that the mutation affects all 3 variants {Fratta, 2013 #1457}.

4.4.3 Summary

Sequence analysis is mainly driven by gene or region candidacy. In this chapter I have used RNA-seq data to generate hypothesis where common variant and post-GWAS analyses had failed. Following up the hypotheses was not possible due to a discovery and a project grant ending that rendered both research aims obsolete.

Chapter 5 Gene expression by anatomy

5.1 Introduction

The spinal regions involved in the lower motor neuron symptoms of ALS are bulbar causing symptoms such as dysphagia, mastication difficulties, slurred speech, and dysarthria; the cervical regions affect upper limb and diaphragmatic function; the thoracic; and lumbar regions affect lower limb function. Each spinal segment has a functional role in ALS which is reflected in pathological changes. This chapter examines whether there are different gene expression profiles that associate with ALS in each region of the spinal cord and whether these inform us about differences in pathology across the spinal cord.

No study of which I am aware has examined changes in gene expression in multiple segments throughout the spinal cord. I hypothesise that each spinal segment will have an inherent normal gene expression profile; specific to the functionality of that region, which is changed in ALS, and which we will see through differential expression analyses comparing cases and controls.

5.2 Methods

5.2.1 Patients

All patients kindly consented to donating their brain and spinal cord to the Medical Research Council's London Brain Bank for Neurodegenerative Disease based at the Institute of Psychiatry, King's College London, or Brains for Dementia Research at King's College London. Please see Table 5-1 for demographic and disease information.

5.2.2 Tissue repository and RNA and DNA isolation

Tissue was flash frozen post-mortem and stored at -80°C as formalin-fixed wax embedded blocks. Tissue blocks of 20mg were taken from available thoracic, cervical, lumbar and medullary regions.

RNA isolation was performed by submersing the 20mg tissue block in 900µl QIAzol lysis reagent within a lysing matrix D provided by MP Biomedicals. The tissue in the lysing matrix was homogenised in a FastPrep 24 for 30 seconds at 4 metres per second Qiagen's RNeasy Universal Kit was used to isolate the RNA. This entailed using gDNA eliminator solution and chloroform for separation of the homogenate into aqueous and organic phases. In several steps the (aqueous) homogenate is mixed with ethanol and washed using Qiagen buffers. This kit uses spin column technology that binds total RNA to a silica membrane, which allows the RNA to be washed using Qiagen buffers and centrifugation. RNA storage temperature was -80°C in RNase-free water.

DNA isolation began by submersing the 25mg tissue block in 80µl PBS and homogenising the tissue using a rotor-stator homogeniser. DNA was isolated using Qiagen's QIAamp DNA Mini and Blood Mini Kit. Proteinase K was added to the homogenate to deactivate protein activity. Similarly, the Qiagen QIAamp protocol uses spin-column technology that allows suspension of DNA in the QIAamp membrane, so the person can wash the DNA using various reagents. DNA was stored at -20°C in Buffer AE (10 mM Tris·Cl; 0.5 mM EDTA; pH 9.0).

5.2.3 RNA and DNA quantification and quality control

RNA quantification was completed using a Life Technologies Qubit 2.0 fluorometer and kit reagents. RNA quality was examined via 260/280 absorbance ratios using a Nanodrop and RNA integrity (RIN) using an Agilent 2100 Bioanalyzer. The RIN algorithm applies electrophoresis and fluorescence to the RNA sample. RNA fragments are separated by molecule size and their fluorescence levels quantified. Degraded samples will show shorter RNA fragments. A RIN score of greater than 7 is the industry standard for successful downstream microarray applications.

However, by using cDNA-mediated Annealing, Selection, Extension, and Ligation (DASL) technology, one can effectively process RNA samples with a RIN number lower than 7. Twelve samples had a RIN score greater than 7, 12 samples had a RIN score between 6 and 7, and 8 samples had a RIN score less than 6. Because certain RNA samples were partially degraded I used Illumina Whole-Genome DASL HT Assay kit for the array expression procedure.

DNA quantification was done using 2µl of the sample solution in a Nanodrop spectrophotometer. DNA quality was inspected using an Agilent 2100 Bioanalyzer.

5.2.4 Whole-Genome Gene Expression using Illumina DASL HT Assays

We used Illumina Human Whole-Genome DASL HT Assay with UDG kit, containing protocol, reagents and BeadChips necessary for the expression analysis, on the Illumina BeadArray platform.

After quantification, the RNA was reverse transcribed using a SUR enzyme, with biotinylated and random nonamer primers. The cDNA was then hybridised to the DASL Assay Pool (DAP) target probes and bound to streptavidin conjugated particles (SA-PMPs).

The supernatant was removed and non-hybridised and mis-hybrised oligos were washed away using UB1. The oligos were then extended and ligated using polymerase to their corresponding downstream-specific oligo (DSO). This created template for PCR.

The oligos then underwent PCR using fluorescently labelled primers, washed, eluted and resuspended into an intermediate plate. The labelled single-stranded PCR product was then isolated and hybridised to the whole-genome expression Illumina BeadChip in an Illumina Hybridisation oven. The BeadChip was then washed using E1BC solution and ethanol, and imaged using an Illumina BeadArray Reader.

5.2.5 Gene expression statistical quality control

Illumina GenomeStudio 2011.1 was utilised to examine and control the quality of RNA expression data.

Normalisation was established by averaging quintiles; standardising the distribution, median and mean probe intensity to the same values for each case. No background subtraction was performed as one platform was used.

Outliers were examined by examining signal averages of control probes by case.

To assess the gene expression data quality and number of genes for each case for each case the housekeeping gene average signal was compared to background noise. The 95th percentile was compared to background noise using as a signal-to-noise ratio to assess quality of expression, and to assess the strength of probe signals. An overall average signal was examined using a box-plot; an average signal > 64,000 was deleted as recommended by Illumina for this gene expression method.

Cluster analyses and related dendrograms were utilised to help to confirm biological replicates and identify any significant outliers. This analysis used the metric 1 – r, *r* being a correlation coefficient of gene expression, for all cases and controls. A scatter plot was performed to examine signal intensities across two samples at a time, with exclusion criteria of $r < 0.99$.

5.2.6 Gene expression statistical analyses

Differential expression analyses were run using Illumina GenomeStudio 2011.1. Both Mann-Whitney and Illumina custom differential tests were used, but the latter is reported here as there

was little incongruence between the two methods. Genes with a false discovery rate <0.01 were excluded. The false discovery rate is a statistic on the proportion of significant results (with $p < 0.05$) that will be false positives. The threshold of 0.01 has been statistically predetermined, similarly to the alpha level of the probability value (p-value). Fold change was calculated by dividing the reference group (controls) average signal with the target group (ALS case) average signal.

Functional annotation clustering and gene ontological analyses of significant differentially expressed gene were analysed using DAVID 6.7[257, 258], and AmiGO. DAVID uses a modified Fisher Exact test to ascribe a metric (EASE score) and p-value to the probability that multiple differentially expressed genes are co-expressing significantly in a cluster. AmiGO was able to identify functional classification of these genes. DAVID also provides an enrichment score, which estimates whether there is over-representation (i.e. enrichment) of target gene cohort in comparison to a reference gene cohort. An enrichment score (for each functional classification) is given to rank the putative importance of the gene-function clusters. It is given p-value from calculating the geometric mean of the EASE values. An annotation cluster enrichment score of 1.3 is equivalent to p-value = 0.05 and therefore annotation clusters with more than $\geq$ 1.3 are mostly reported, unless stated otherwise. Benjamini and Hochberg corrections are reported which corrects p-values to correct for multiple testing. I have not excluded cluster GO terms that were as it the statistic is conservative and enrichment scores are my main statistic of interest in identifying relevant genes that may be involved in ALS.

5.2.7 Bioinformatics interaction analyses

Genes found to show significant differences in expression and those highly enriched in the functional annotation cluster analysis were examined using the protein-protein network tool STRING 9.05[259] and the gene-gene and gene-protein network tool GeneMANIA[260]. Networks showing some relevance with previous ALS research are reported as I am interested in their candidacy. Follow-up analyses exploring the association and interactions between genes were performed using BioGraph[261].

5.3 Results

5.3.1 Patient Characteristics

Please see Table 5-1 for demographic and disease information for cases and controls

	Sex	Age at death (yrs)	Post Mortem Delay (hrs)	AOO (yrs)	ALS Type
ALS_1	F	75	38	72.4	Limb onset
ALS_2	F	80	37	65.4	Limb onset
ALS_3	F	65	14	61.3	Limb onset
ALS_4	F	63	25	58.6	Limb onset
ALS_5	M	50	26	46.5	Limb onset
ALS_6	F	51	33	Missing	ALS-FTD
CTRL_1	F	54	31	-	-
CTRL_2	M	89	41	-	-
CTRL_3	F	21	21	-	-

Table 5-1. Case and controls demographic information with disease information for cases

5.3.2 Differential analyses comparing cases with control using all spinal segments

Comparing cases with controls, 167 genes were selected based on an expression difference p-value of < 0.05, and with the exclusion of genes with a false-discovery detection-threshold ratio of < 0.01 (see Appendix Table A8). None of these genes had any reported involvement in ALS pathology using a literature search.

The 167 genes were analysed using DAVID functional annotation clustering. Clusters with an enrichment score of ≥1.3 were selected and are displayed in Table 5-2 (function: Glycosylation & Transmembrane activity), 5-3 (function: Symporter activity), and 5-4 (function: Glycan metabolic process). Table 5-5 (function: Calcium Binding) is also included as its function is relevant to known ALS pathology and gene expression research. The list of genes, and corresponding fold-change, that belong to each functional classification can be found in the appendix table A9.

Gene *MSRA*, implicated in glycan metabolic processes, showed an interesting network connection with *KIF3A* (Figure 5-1), as did *PLCD3* with *ITPR1* (Figure 5-2). These genes were selected as *KIF3A* is a part of the kinese II complex with ALS gene *KIFAP3*, and *ITPR1* is an ALS gene whose protein product interacts with TDP-43.

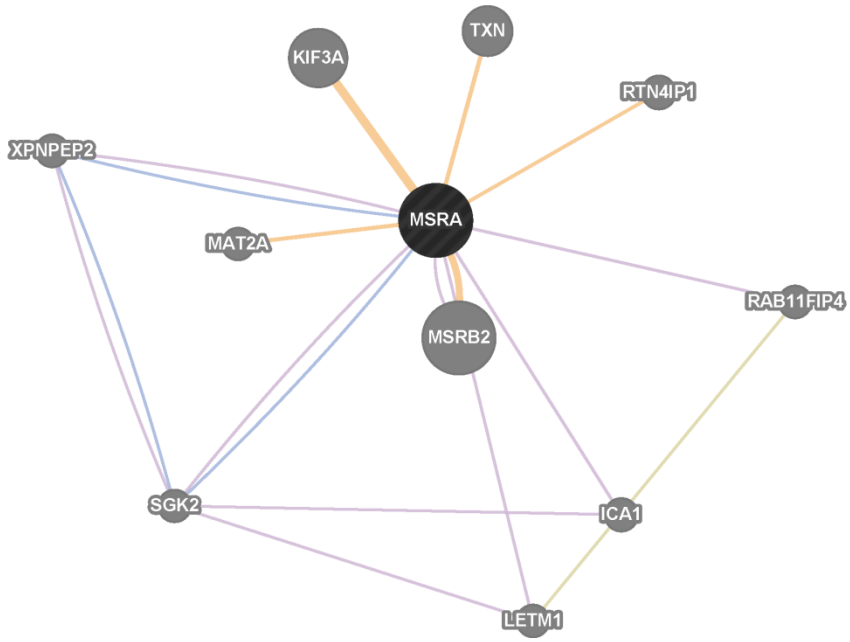


Figure 5-1. Gene Protein network map of *MSRA* using GeneMANIA

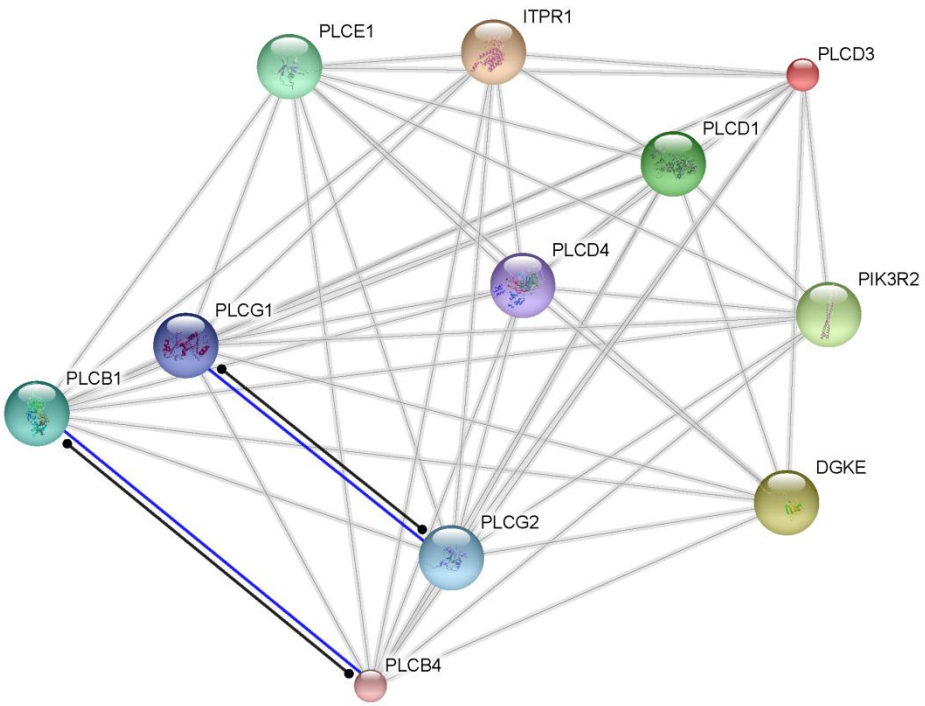


Figure 5-2. Protein-Protein network map of *PLCD3* using STRING

Enrichment Score: 1.68

Glycosylation & Transmembrane activity

Term	*Gene count*	*% of Total Genes*	*EASE P-value*	*Fold Enrichment*	*Benjamini Correction*
glycosylation site: N-linked (GlcNAc...)	48	30.77	>0.01	1.49	0.70
Glycoprotein	48	30.77	0.01	1.44	0.73
GO:0031224~intrinsic to membrane	56	35.90	0.02	1.25	0.97
Transmembrane region	50	32.05	0.03	1.31	0.88
Transmembrane	50	32.05	0.03	1.30	0.97
topological domain: Cytoplasmic	36	23.08	0.04	1.37	0.90
GO:0016021~integral to membrane	53	33.97	0.04	1.23	0.97
topological domain: Extracellular	30	19.23	0.04	1.42	0.88
Membrane	59	37.82	0.05	1.22	0.91

Table 5-2. Most enriched functional cluster: involved in glycosylation and transmembrane activity

Symporter activity

Enrichment Score: 1.41

Term	*Gene count*	*% of Total Genes*	*EASE P-value*	*Fold Enrichment*	*Benjamini Correction*
GO:0015295~solute:hydrogen symporter activity	3	1.92	0.01	15.90	0.91
GO:0015294~solute:cation symporter activity	4	2.56	0.03	5.58	0.97
Symport	4	2.56	0.05	4.65	0.89
GO:0015293~symporter activity	4	2.56	0.08	3.87	0.99

Table 5-3. Second most enriched functional cluster: involved in symporter activity

Enrichment Score: 1.37

Term	*Gene count*	*% of Total Genes*	*EASE P-value*	*Fold Enrichment*	*Benjamini Correction*
GO:0006790~sulfur metabolic process	6	3.85	>0.01	6.66	0.85
GO:0006029~proteoglycan metabolic process	3	1.92	0.04	8.90	1.00
GO:0005976~polysaccharide metabolic process	4	2.56	0.06	4.60	1.00
GO:0030203~glycosaminoglycan metabolic process	3	1.92	0.07	6.96	1.00
GO:0006022~aminoglycan metabolic process	3	1.92	0.09	5.89	1.00
GO:0009100~glycoprotein metabolic process	4	2.56	0.21	2.53	1.00

Glycan metabolic processes

Table 5-4. Third most enriched functional cluster: involved in metabolic processes

Enrichment Score: 1.13

Calcium Binding

Term	*Gene count*	*% of Total Genes*	*EASE P-value*	*Fold Enrichment*	*Benjamini Correction*
calcium-binding region:1; low affinity	3	1.92	0.01	21.38	0.87
calcium-binding region:2; high affinity	3	1.92	0.01	18.33	0.84
domain:EF-hand 2	5	3.21	0.05	3.52	0.91
domain:EF-hand 1	5	3.21	0.05	3.52	0.91
EF hand	3	1.92	0.08	6.45	0.91
IPR018247:EF-HAND 1	5	3.21	0.08	3.02	1.00
IPR011992:EF-Hand type	5	3.21	0.10	2.78	1.00
GO:0005509~calcium ion binding	11	7.05	0.15	1.59	1.00
Calcium	10	6.41	0.17	1.61	0.97
calcium binding	3	1.92	0.19	3.76	0.97
IPR018249:EF-HAND 2	3	1.92	0.49	1.83	1.00

Table 5-5. Fourth most enriched functional cluster: involved in calcium binding.

5.3.3 Differential analyses comparing cases with controls by spinal segments

5.3.3.1 Medulla

Comparing cases with controls using just medulla samples identified 115 differentially expressed genes, which were analysed with DAVID functional annotation clustering. The main functional cluster identified was regulation of insulin, peptides and hormone secretion, with an enrichment score of 1.22 (Table 5-6). Other clusters had poor enrichment scores, with functions involved in cell junctions (enrichment score: 1.06) and embryonic organ development (enrichment score: 0.90). See appendix Table A10 for a list of differentially expressed genes involved in the top functional cluster.

5.3.3.2 Cervical

Comparing cases with controls identified only 12 differentially expressed genes in the cervical samples, which were analysed with DAVID functional annotation clustering. The main functional cluster identified was regulation of muscle contraction, response to external stimulus & apoptosis, with an enrichment score of 1.20 (Table 5-7 and appendix table A11). The gene *ADA*, which differentially expressed in this category, was predicted to interact with ALS2 and known to interact with *DPP4* (Figure 5-3). No other cluster had a reasonably high enrichment score or showed relevance in terms of function with known ALS pathology.

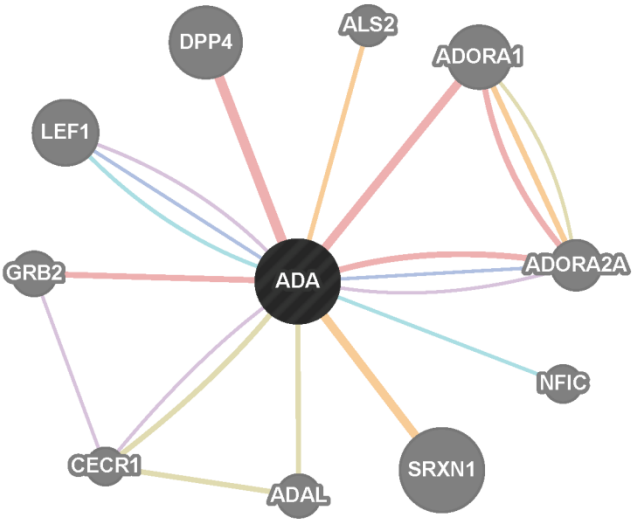


Figure 5-3. Gene Protein network map of *ADA* using GeneMANIA.

Purple line indicates a co-expression, pink line physical interactions, orange line predicted interaction, light blue line shared pathway, dark blue line co-localisation, and a beige line shared protein domains.

5.3.3.3 Thoracic

Comparing cases with controls using thoracic cord samples identified 20 differentially expressed genes, which were analysed with DAVID functional annotation clustering. The main functional cluster identified was involved in conjugation pathways & catabolic processes, with an enrichment score of 0.95 (Table 5-8 and appendix table A12). The gene *UFD1L*, which differentially expressed in this category, binds with *VCP* to form a protein complex. No other cluster had a reasonably high enrichment score or showed relevance in terms of function with known ALS pathology.

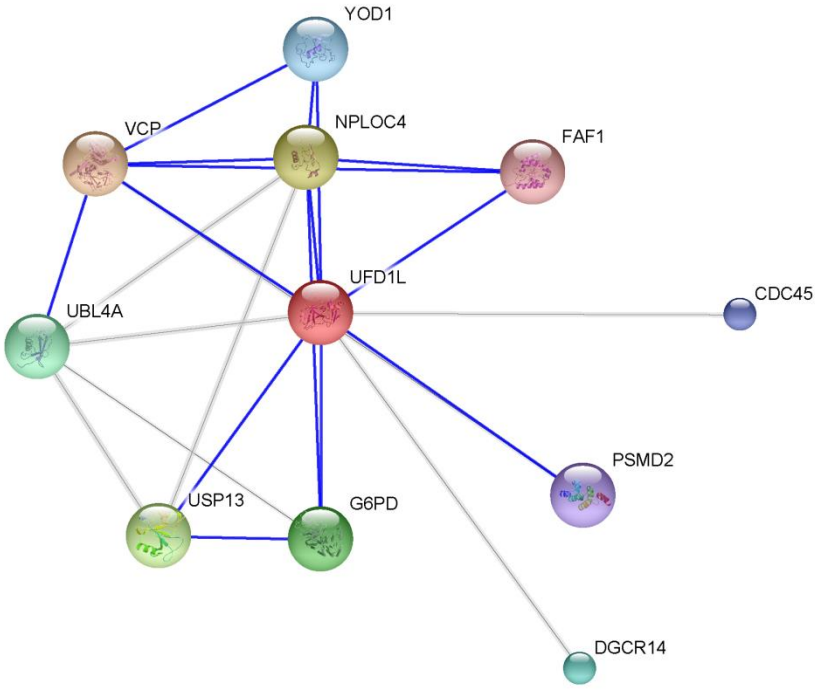


Figure 5-4. Protein-Protein network map of *UFD1L* using STRING

5.3.3.4 Lumbar

Comparing cases with controls using lumbar samples identified 29 differentially expressed genes, which were analysed with DAVID functional annotation clustering. The main functional cluster identified was involved in flavoproteins and oxidoreductase function, having the highest enrichment score of 1.68 across all spinal segments (Table 5-9 and appendix table A13). The gene *CYB5R1*, which showed differential expression in this cluster, binds with *UBQLN4* which, similarly to *VCP* and *UBQLN2*, regulates proteosomal protein catabolic processes (Figure 5-5). A second significant functional cluster implicated pleckstrin homology domains with an enrichment score of 1.63 (see Table 5-10). No other cluster had a reasonable enrichment score or showed relevance in terms of function with known ALS pathology.

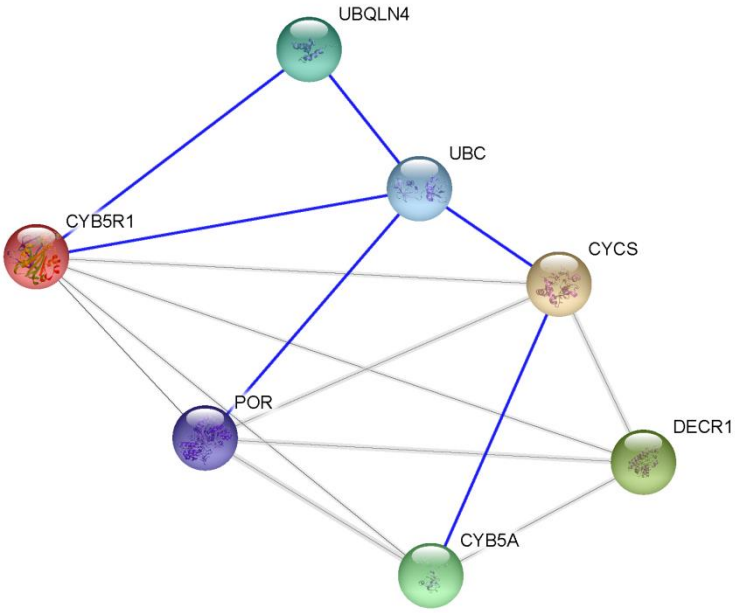


Figure 5-5. Protein-Protein network map of *CYB5R1* using STRING

	Enrichment Score: 1.22					
	Term	**Gene count**	**% of Total Genes**	**EASE P-value**	**Fold Enrichment**	**Benjamini Correction**
Regulation of insulin, peptide and hormone secretion	**GO:0032024~positive regulation of insulin secretion**	**3**	**2.70**	**>0.01**	**27.84**	**0.90**
	GO:0002793~positive regulation of peptide secretion	**3**	**2.70**	**0.01**	**21.78**	**0.84**
	GO:0046887~positive regulation of hormone secretion	**3**	**2.70**	**0.02**	**13.92**	**0.95**
	GO:0050796~regulation of insulin secretion	**3**	**2.70**	**0.02**	**12.53**	**0.93**
	GO:0002791~regulation of peptide secretion	**3**	**2.70**	**0.03**	**10.89**	**0.94**
	GO:0046883~regulation of hormone secretion	**3**	**2.70**	**0.06**	**7.59**	**0.93**
	GO:0051046~regulation of secretion	**4**	**3.60**	**0.12**	**3.31**	**0.98**
	GO:0051047~positive regulation of secretion	**3**	**2.70**	**0.14**	**4.60**	**0.99**
	GO:0051050~positive regulation of transport	**3**	**2.70**	**0.38**	**2.25**	**1.00**
	GO:0010033~response to organic substance	**6**	**5.41**	**0.42**	**1.39**	**1.00**
	GO:0060341~regulation of cellular localization	**3**	**2.70**	**0.43**	**2.02**	**1.00**

Table 5-6. Most enriched medulla functional cluster involved in regulation of insulin, peptides and hormone secretion

Regulation of muscle contraction, response to external stimulus & apoptosis

Enrichment Score: 1.20

Term	*Gene count*	*% of Total Genes*	*EASE P-value*	*Fold Enrichment*	*Benjamini Correction*
GO:0006937~regulation of muscle contraction	3	27.27	>0.01	62.63	0.28
GO:0010035~response to inorganic substance	3	27.27	0.01	22.00	0.73
GO:0010648~negative regulation of cell communication	3	27.27	0.01	18.18	0.72
GO:0044057~regulation of system process	3	27.27	0.01	14.59	0.77
GO:0042127~regulation of cell proliferation	3	27.27	0.07	5.73	1.00
GO:0042981~regulation of apoptosis	3	27.27	0.08	5.61	1.00
GO:0043067~regulation of programmed cell death	3	27.27	0.08	5.55	0.99
GO:0010941~regulation of cell death	3	27.27	0.08	5.53	0.99
GO:0046872~metal ion binding	5	45.45	0.15	1.96	1.00
GO:0043169~cation binding	5	45.45	0.16	1.94	1.00
GO:0043167~ion binding	5	45.45	0.16	1.91	0.98
GO:0046914~transition metal ion binding	3	27.27	0.46	1.75	1.00
Polymorphism	7	63.64	0.63	1.06	1.00
sequence variant	7	63.64	0.70	1.01	1.00

Table 5-7. Most enriched cervical functional cluster involved in regulation of muscle contraction, response to external stimulus & apoptosis

Conjugation pathways & catabolic processes	**Enrichment Score: 0.95**					
	Term	***Gene count***	***% of Total Genes***	***EASE P-value***	***Fold Enrichment***	***Benjamini Correction***
	ubl conjugation pathway	3	15	0.07	6.30	0.78
	GO:0019941~modification-dependent protein catabolic process	3	15	0.09	5.44	1.00
	GO:0043632~modification-dependent macromolecule catabolic process	3	15	0.09	5.44	1.00
	GO:0051603~proteolysis involved in cellular protein catabolic process	3	15	0.10	5.20	1.00
	GO:0044257~cellular protein catabolic process	3	15	0.10	5.18	1.00
	GO:0030163~protein catabolic process	3	15	0.10	5.02	1.00
	GO:0044265~cellular macromolecule catabolic process	3	15	0.13	4.31	1.00
	GO:0009057~macromolecule catabolic process	3	15	0.15	4.00	1.00
	GO:0006508~proteolysis	3	15	0.24	2.96	1.00

Table 5-8. Most enriched thoracic functional cluster involved in conjugation pathways & catabolic processes

	Enrichment Score: 1.68					
Flavoproteins and oxidoreductase	**Term**	**Gene count**	**% of Total Genes**	**EASE P-value**	**Fold Enrichment**	**Benjamini Correction**
	Flavoprotein	**3**	**1.57**	**0.01**	**19.97**	**0.50**
	FAD	**3**	**1.57**	**0.01**	**18.75**	**0.32**
	Oxidoreductase	**4**	**2.09**	**0.04**	**5.07**	**0.63**
	GO:0055114~oxidation reduction	**4**	**2.09**	**0.05**	**4.46**	**1.00**

Table 5-9. Most enriched lumbar functional cluster involved in flavoproteins and oxidoreductase

Pleckstrin homology	Enrichment Score: 1.63					
	Term	**Gene count**	**% of Total Genes**	**EASE P-value**	**Fold Enrichment**	**Benjamini Correction**
	IPR011993:Pleckstrin homology-type	4	2.09	0.01	8.80	0.50
	SM00233:PH	3	1.57	0.02	10.93	0.24
	IPR001849:Pleckstrin homology	3	1.57	0.06	7.22	0.90

Table 5-10. Second most enriched lumbar functional cluster involved in pleckstrin homology domains

5.4 Discussion

Differential expression analyses identified significant changes in gene expression in spinal regions collectively and in each anatomical spinal region separately, when comparing ALS cases and non-ALS controls.

5.4.1 Functional categories of differentially expressing genes combining all anatomical regions

5.4.1.1 Glycosylation and transmembrane activity

The most enriched annotated function identified when comparing all anatomical regions across cases and controls was for Nitrogen(N)-linked glycosylation and transmembrane activity. Glycosylation refers to the formation of glycoproteins. These are multifunctional ubiquitous proteins. One major function of such proteins is to act as receptors on cell membranes, which plays an important role with transmembrane activity. Aberrant, possibly reduced, glycosylation of an excitatory amino-acid transporter was implicated in ALS as a possible mechanism of disease, as a consequence of a point mutation in *EAAT2*[262]. Increased levels of sialylated glycans and lowered levels of fucosylated glycans, examined in N-glycans, were found comparing ALS cases against controls[263]. Recently a glycoprotein *GPNMB* was found to be increased in mutant-*SOD1* (*SOD1*^{MT}) mice and ALS patients' spinal cords, consistent with the findings here (see Appendix Table A8)[264].

For alternative glycans, Oxygen(O)-linked glycosylation in the tail domain of neurofilament protein M has been tentatively implicated in ALS[265]. O-linked glycosylation with β-N-acetylglucosamine residues have been found to be reduced in spinal cord motor neurons in *SOD1*^{MT} mice[266].

Transmembrane activity refers to proteins which pass through the cell membrane in either direction, and to functions in guiding proteins in or out of the cell. Prensenilin-1 (PSEN1) is a

transmembrane protein and also implicated in apoptosis, another process associated strongly with ALS pathology[191]. PSEN-1 was associated with ALS in 2000, but it has been difficult to replicate this finding in ALS excluding FTD. Since this discovery, several genes have been implicated in dysfunctional membrane trafficking, including *VAPB* in a family linkage study[267], *DPP6* identified in an Irish Genome Wide Association Study (GWAS)[114], *FIG4* via a nonsynonymous mutation[234], *SEMA6A* in a GWAS[218], the transmembrane domain of *SIGMAR1* in a familial homozygosity mapping study[268], and *CRIM1* from a resequencing analysis.

GPNMB was also found to be differentially expressed in this category echoing previous findings using *SOD1*(G93A) mice, suggesting that it may be a neuroprotective factor[264].

5.4.1.2 Symporter activity

Symporter activity describes the action of symports, which are membrane proteins that transport two or more different molecules through the membrane simultaneously. This action has not been implicated in ALS pathology so far. Vesicle-associated membrane protein-associated protein C (*VAMP*) and the *FIG4* gene, both found to cause ALS pathology when mutated, function as membrane regulators.

5.4.1.3 Glycan metabolic processes

The glycan metabolic processes cluster is related to glycan and glycoprotein related activity, the most enriched functional category in this section (please refer to 5.1.1). A literature search into sulphur, proteoglycan, polysaccharide, glycosaminoglycan, aminoglycan and glycoprotein metabolic processes in ALS revealed little reported association between glycan related metabolic processes and ALS.

One gene identified in this category was peptide methionine sulfoxide reductase *MSRA*. It functions as a repair protein in response to oxidative damage, and showed several interesting network associations. There is a strong predicted functional relationship between *MSRA* and *KIF3A*. *KIFAP3*, a gene previously found to affect survival in ALS[218, 269], acts as an important mediator in *KIF3A* function. Furthermore *KIF3A* proteins have been found to be under-expressed in ALS motor cortex[270]. Figure 5-6 displays multiple routes in which *MSRA* could affect or be affected by *KIFAP3*.

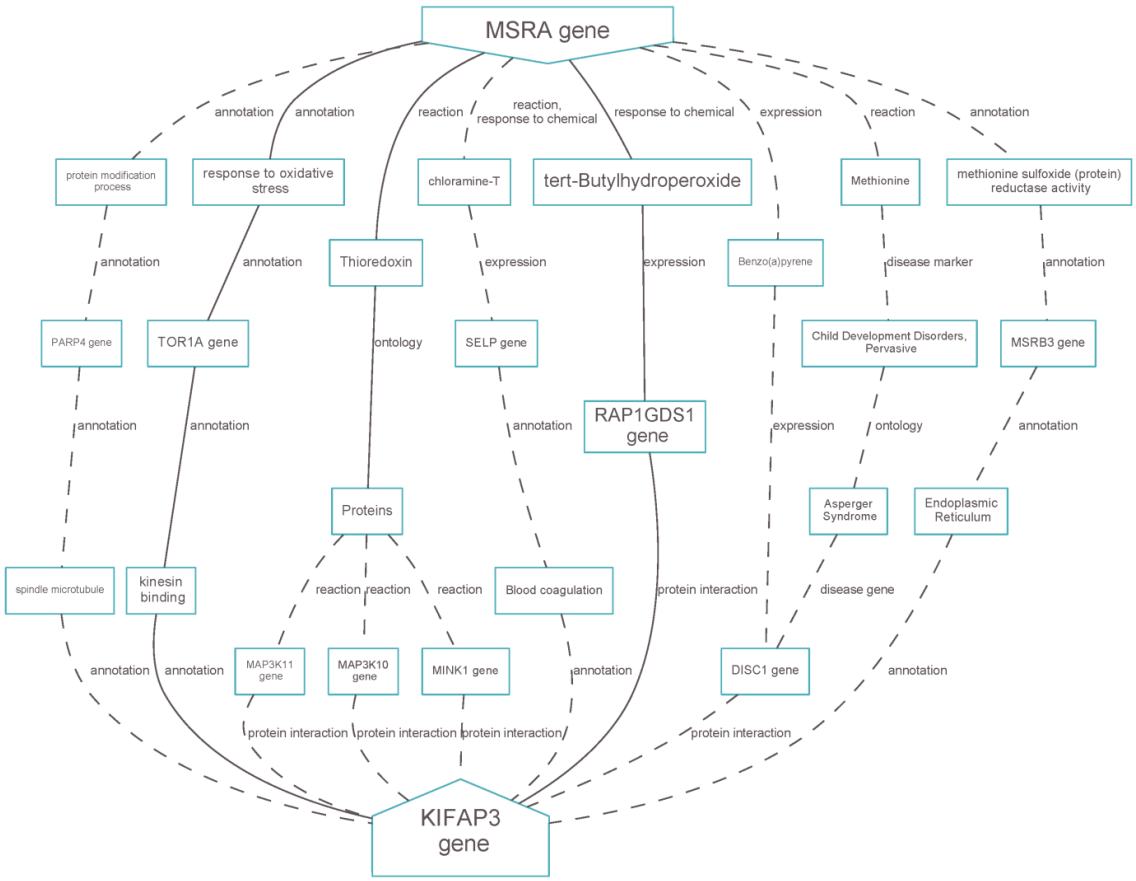


Figure 5-6. Functional relationships between *MSRA* and *KIFAP3*. Solid lines represent well-supported associations; dotted lines represent less-supported associations.

5.4.1.4 Calcium binding

Calcium-binding has long been implicated in ALS pathology[271]. Intracellular calcium levels are raised in ALS mice and patients[272], especially linked to those with *SOD1* mutations[273]. Intracellular calcium activity has also been linked to *SOD1* based putative pathological mechanisms; glutamatergic excitotoxicity[274] and free radical propagation[275], exacerbating and being exacerbated by both.

In the current analysis, *PLCD3* was found to differentially express between cases and controls when using all spinal samples. *PLCD3* is closely related to *PLCD1*, both belonging to phospholipase C delta protein family. *PLCD1* has been implicated in ALS through a genome-wide microsatellite association analysis[214], and there is evidence of significant over-expression of *PLCD1* RNA and protein[215]. *PLCD3* also interacts with *ITPR1*, which when knocked down assists autophagosomal clearance of TDP-43[276], a hallmark of pathology in ALS. *ITPR1* is likely to be important for nucleoplasmic shuttling and proteostasis of TDP-43.

5.4.2 Functional categories of differentially expressing genes for each specific anatomical section

5.4.2.1 Medulla - Regulation of insulin, peptide and hormone secretion

Abnormal secretion of pancreatic insulin in ALS was reported in 1972[277]. Insulin secretion was later believed by some to be a secondary effect and modify survival[278], and it led to trials in ALS using insulin growth factors (IGF)[279]. Later, IGF-1 was found to be reduced in the ventral horn of ALS[280] and was linked to *SOD1* pathology[281]. Recently insulin was found to reduce the aggregation of TDP-43 in human cells[282]

Thyrotropin-releasing hormone was seen as a hopeful candidate for clinical trials as it had been previously seen to be decreased in the anterior horn of ALS cases[283]. Unfortunately unexpected

negative pulmonary consequences upon the patient impeded this research until it was found through several trials not to alter survival[284-286].

5.4.2.2 Cervical - Regulation of muscle contraction, response to external stimulus & apoptosis

The gene *ADA* showed differential expression in cervical spinal cord. It aids immunodevelopment as well as maintenance, and its deficiency causes severe combined immunodeficiency disease. GeneMANIA reported a predicted relationship with ALS2, a gene involved in endosomal trafficking and known to cause juvenile-onset ALS.[287, 288] The network also revealed a known interaction between *ADA* and *DPP4*. *DPP4* shares significant sequence homology with *DPP6* and belongs to the same dipeptidyl aminopeptidase-like protein family. *DPP6* was first implicated in ALS through a genome-wide association study. The functional relationship of how *ADA* and *DPP6* may interact through *DPP4* is visualised in Figure 5-7.

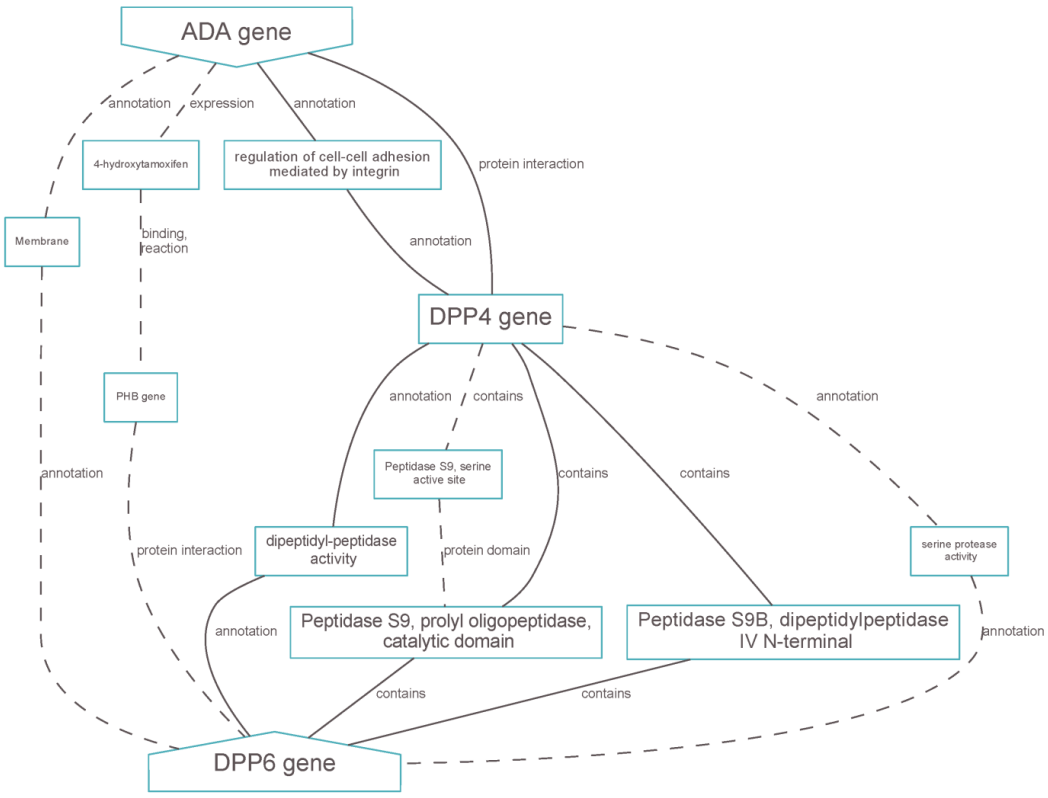


Figure 5-7. Functional relationships between *ADA* and *DPP6*, intermediated by *DPP4*.
Solid lines represent well-supported associations; dotted lines represent less-supported associations.

5.4.2.3 Thoracic - Conjugation pathway & catabolic processes

One of the three genes in this functional group was *UFD1L*. The *UFD1L* protein is known to degrade ubiquitinated proteins through forming a complex with *NPL4* and *VCP*. *VCP* is involved in vesicle transport, membrane fusion and protein degradation. The *VCP* gene is reported to harbour point mutations that cause ALS[289].

Pathological remnants of a faulty ubiquitinated recycle process(es) are considerably present in ALS post-mortem tissue. We know point mutations in a gene *UBQLN2* cause ALS[290], resulting in the interruption of protein degradation regulation, using conjugated markers to identify targeted proteins. *UBQLN2* encodes for ubiquitin-like protein ubiquilin, which is a very salient biomarker in ALS tissue. It is an important finding for ALS as it is the first genetic mutation that is directly linked to this biomarker, and this biomarker is present regardless of what genetic mutations the patient has. Furthermore, it implicates a dysfunction of ubiquilin-based recycling processes as a direct cause, instead of consequence, of ALS. In this analysis it is supported by the possible dysfunction caused by *VCP* and *UFD1L* under-expression.

5.4.2.4 Lumbar - Flavoproteins and oxidoreductase

The most ALS relevant gene from this group, *CYB5R1*, is involved in desaturation and elongation of fatty acids, cholesterol biosynthesis, drug metabolism, and in erythrocyte reduction. It binds with *UBQLN4*, which also aids regulation of protein degradation, which in turn interacts with *UBQLN2*[290], sharing considerable sequence homology (Figure 5-8). Furthermore *UBQLN4* interacts with *ATXN1*, which causes spinocerebellar ataxia; a disease with known overlap with ALS[97].

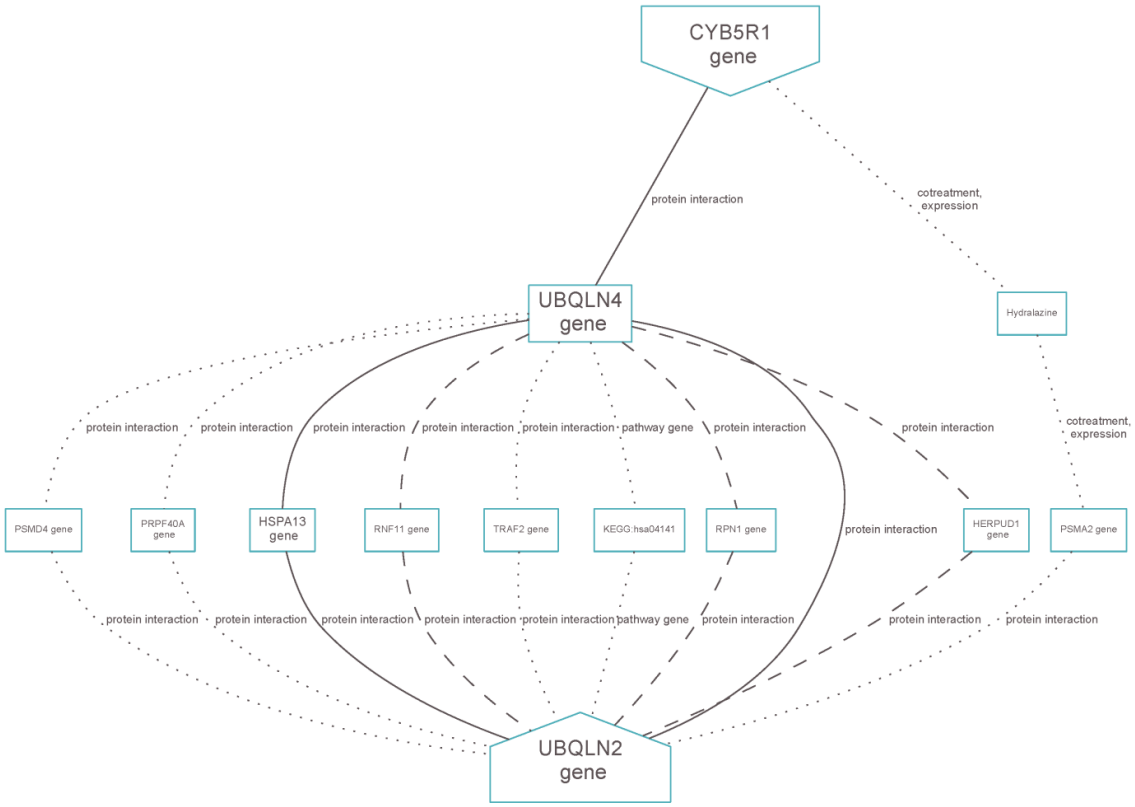


Figure 5-8. Functional relationships between *CYB5R1* and *UBQLN2*, intermediated by *DPP4*. Solid lines represent well-supported associations; dotted lines represent less-supported associations.

5.4.2.5 Lumbar - Pleckstrin homology

Pleckstrin homology domains occur in a wide range of proteins, modulating intracellular signalling and membrane protein interaction. ALS2 has a pleckstrin homology domain alongside an N-terminal RCC1-like domain, Dbl homology, and a C-terminal VPS9 domain. Pleckstrin homology domain-containing protein, family h, member 2 (*PLEKHH2*) was found in this functional group to differentially express between cases and controls. *PLEKHH2* and *PON1* are believed to be implicated in microvascular complications in diabetes 5, although this is not yet replicated. For ALS, *PON1* is believed to protect against toxins that may augment lipid oxidation pathology. There

are point mutations in this gene associated with ALS and an ALS-associated haplotype thought to decrease gene expression[291-294].

5.4.3 Methodology criticism

Like all gene expression analyses, it is difficult to establish how much gene activity is a consequence of responses to the disease itself, how much is directly involved in ALS pathology. For this chapter there is an additional dimension, of how much of the genes identified are a consequence of the chosen spinal segment and the extent of which it is affected. It follows that, for example, a majority of upper-limb onset patients will have cervical segments heavily affected whereas lumbar and medulla may be less so. Furthermore we are comparing across pathological severity because we have a mixed number of patients with different onset locations, which will cause a different pattern of disease spread. Therefore knowing the spread and pathology of the disease is an important dimension, which is what is examined in the following paragraph. Through direct comparison with findings from the following paragraph, we are able to identify those genes showing differential expression as consequence of spread and pathology, and those genes which are important for ALS at a specific spinal anatomical region.

Screening the patients used in this chapter for genetic mutation would have been insightful as we may have seen, or rather been able to ascertain, whether the gene expression profiles identified were related to a particular mutation. It may be that genetic mutations create differences in gene expression profiles between patients.

The method utilised GenomeStudio as its main statistical program. There has been criticism against using proprietary software in gene expression studies. As a comparison with GenomeStudio, the R package Lumi {Du, 2008 #1458} integrates additional quality control steps and allows the user more freedom in selecting alternative normalisation, variance stabilisation, and background correction methods. Additional quality control plots, differential expression, and

gene annotation methods can be achieved using non-proprietary software, which may have led to clearer and alternative results.

5.4.4 Summary

The functional categories statistically attributed to groups of differentially expressing genes replicated some well-established pathological pathways and identified new gene candidates.

Transmembrane dysfunction is a disease mechanism with much pathological and genetic evidence in ALS; however I also find here that symporter activity may play a major functional role. Symporter activity has been implicated before in ALS, especially with regards to the gene *SLC1A2*. The involvement of calcium-binding proteins has had strong support from ALS research since the early 1990s, and this is also suggested by the results presented here. I also found glycans and glycosylation activity in the gene expression profiles, both of which have increasing evidence for involvement in ALS through pathology research.

The insulin pathway has been somewhat implicated through insulin-like growth factor 1 (IGF-1), as have flavoproteins, both of which have moderate support in the results presented here. Muscle contraction, apoptosis and oxidoreductase have a significant role in ALS, and this was also replicated in this chapter. We found conjugation pathways, catabolic process, and pleckstrin homology to be of importance in ALS pathogenesis based on differential expression of RNA but these have had very little evidence for involvement in ALS previously. This may be a consequence as little research has been performed on thoracic segments of the spinal cord.

Chapter 6 Gene expression by spread and pathology

6.1 Introduction

There is evidence to support the idea that ALS spreads sequentially through each anatomical segment of the nervous system from an initial onset point. Yet ALS has a definitive genetic component, and it is unclear how this component interacts with the spread of disease. My objective is to examine how ALS genetics plays a role in each segment of the spinal cord as it becomes sequentially drawn in to the disease, and to examine ALS genetics at different levels of pathological severity.

By using differential expression analyses I assess whether gene activity at different disease stages and pathological severity categories alter as a consequence of having ALS. Although one cannot draw inferences about causation or response, it may be possible to hypothesise whether a change is primary or secondary based on the pattern of expression changes.

Clinical observation is largely consistent with the idea that ALS, in most cases, spreads contiguously through the spinal cord and brain stem[156]. Motor neural density of a spinal cord region is decreased by the extent of time ALS has been present in that region. It follows that this density decreases as one moves away from the epicentre of the disease[150, 295]. Similarly, the presence of ALS inclusions, such as the hallmark ubiquitin-positive aggregates, is again relative to the extent of time ALS has been present in an anatomical region. Proteins with conformational abnormalities have been shown to be toxic for motor neurons, and it has been hypothesised that spread of ALS disease is via inter-neural prion-like propagation. An alternative hypothesis is that protein abnormalities and motor neuron death are due to the spread of a toxin with an external origin.

In this chapter I estimate contiguous disease spread through anatomy using clinical symptomatological data alongside estimations of motor neural density that used haematoxylin and eosin (H&E) stains. I then examine changes in gene expression in these spinal regions parallel with the inferred spread of disease, to assess changes in gene expression present at

each location. I also examine whether any differential expression genes show a dose-dependent response to the clinical description of disease burden at each segment.

In the second analysis I ignore the spread of disease and examine directly the severity of pathology. This, I have taken as an indicator of the time that ALS pathology has been present in that spinal region. With four categories of rating pathological severity, I assess whether each implicates different genes and has a different gene expression profile. This will help elucidate if there is involvement of different genes in ALS as a consequence of severity and the time ALS pathology has been present in that anatomical segment.

6.2 Methods

6.2.1 Patients

All patients kindly consented to donating their brain and spinal cord to the Medical Research Council's London Brain Bank for Neurodegenerative Disease based at the Institute of Psychiatry, King's College London, or Brains for Dementia Research at King's College London. Please see Table 6-1 for demographic and disease information.

6.2.2 Tissue repository and RNA and DNA isolation

Tissue was flash frozen post-mortem and stored at -80°C as formalin-fixed wax embedded blocks. Tissue blocks of 20mg were taken from available thoracic, cervical, lumbar and medullary regions.

RNA isolation was performed by submersing the 20mg tissue block in 900µl QIAzol lysis reagent within a lysing matrix D provided by MP Biomedicals. The tissue in the lysing matrix was homogenised in a FastPrep 24 for 30 seconds at 4 metres per second Qiagen's RNeasy Universal

Kit was used to isolate the RNA. This entailed using gDNA eliminator solution and chloroform for separation of the homogenate into aqueous and organic phases. In several steps the (aqueous) homogenate is mixed with ethanol and washed using Qiagen buffers. This kit uses spin column technology that binds total RNA to a silica membrane, which allows the RNA to be washed using Qiagen buffers and centrifugation. RNA storage temperature was -80°C in RNase-free water.

DNA isolation began by submersing the 25mg tissue block in 80µl PBS and homogenising the tissue using a rotor-stator homogeniser. DNA was isolated using Qiagen's QIAamp DNA Mini and Blood Mini Kit. Proteinase K was added to the homogenate to deactivate protein activity. Similarly, the Qiagen QIAamp protocol uses spin-column technology that allows suspension of DNA in the QIAamp membrane, so the person can wash the DNA using various reagents. DNA was stored at -20°C in Buffer AE (10 mM Tris·Cl; 0.5 mM EDTA; pH 9.0)

6.2.3 RNA and DNA quantification and quality control

RNA quantification was completed using a Life Technologies Qubit 2.0 fluorometer and kit reagents. RNA quality was examined via 260/280 absorbance ratios using a Nanodrop and RNA integrity (RIN) using an Agilent 2100 Bioanalyzer. The RIN algorithm applies electrophoresis and fluorescence to the RNA sample. RNA fragments are separated by molecule size and their fluorescence levels quantified. Degraded samples will show shorter RNA fragments. A RIN score of greater than 7 is the industry standard for successful downstream microarray applications. However, by using cDNA-mediated Annealing, Selection, Extension, and Ligation (DASL) technology, one can effectively process RNA samples with a RIN number lower than 7. Twelve samples had a RIN score greater than 7, 12 samples had a RIN score between 6 and 7, and 8 samples had a RIN score less than 6. Because certain RNA samples were partially degraded I used Illumina Whole-Genome DASL HT Assay kit for the array expression procedure.

DNA quantification was done using 2µl of the sample solution in a Nanodrop spectrophotometer. DNA quality was inspected using an Agilent 2100 Bioanalyzer.

6.2.4 Whole-Genome Gene Expression using Illumina's DASL HT Assays

We used Illumina Human Whole-Genome DASL HT Assay with UDG kit, containing protocol, reagents and BeadChips necessary for the expression analysis, on the Illumina BeadArray platform.

After quantification, the RNA was reverse transcribed using a SUR enzyme, with biotinylated and random nonamer primers. The cDNA was then hybridised to the DASL Assay Pool (DAP) target probes and bound to streptavidin conjugated particles (SA-PMPs).

The supernatant was removed and non-hybridised and mis-hybrised oligos were washed away using UB1. The oligos were then extended and ligated using polymerase to their corresponding downstream-specific oligo (DSO). This created template for PCR.

The oligos then underwent PCR using fluorescently labelled primers, washed, eluted and resuspended into an intermediate plate. The labelled single-stranded PCR product was then isolated and hybridised to the whole-genome expression Illumina BeadChip in an Illumina Hybridisation oven. The BeadChip was then washed using E1BC solution and ethanol, and imaged using an Illumina BeadArray Reader.

6.2.5 Gene expression statistical quality control

Illumina GenomeStudio 2011.1 was utilised to examine and control the quality of RNA expression data.

Normalisation was established by averaging quintiles; standardising the distribution, median and mean probe intensity to the same values for each case. No background subtraction was performed as one platform was used.

Outliers were examined by examining signal averages of control probes by case.

To assess the gene expression data quality and number of genes for each case for each case the housekeeping gene average signal was compared to background noise. The 95th percentile was compared to background noise using as a signal-to-noise ratio to assess quality of expression, and to assess the strength of probe signals. An overall average signal was examined using a box-plot; an average signal > 64,000 was deleted as recommended by Illumina for this gene expression method.

Cluster analyses and related dendrograms were utilised to help to confirm biological replicates and identify any significant outliers. This analysis used the metric 1 – r, *r* being a correlation coefficient of gene expression, for all cases and controls.. A scatter plot was performed to examine signal intensities across two samples at a time, with exclusion criteria of r < 0.99.

6.2.6 Gene expression statistical analyses

Differential expression analyses were run using Illumina GenomeStudio 2011.1. Both Mann-Whitney and Illumina custom differential tests were used, but the latter is reported here as there was little incongruence between the two methods. Genes with a false discovery rate <0.01 were excluded. The false discovery rate is a statistic on the proportion of significant results (with p < 0.05) that will be false positives. The threshold of 0.01 has been statistically predetermined, similarly to the alpha level of the probability value (p-value).

Functional annotation clustering and gene ontological analyses of significant differentially expressed gene were analysed using DAVID 6.7[257, 258] and AmiGO. DAVID uses a modified

Fisher Exact test to ascribe a metric (EASE score) and p-value to the probability that multiple differentially expressed genes are co-expressing significantly in a cluster. AmiGO was able to clarify functional classification of these genes. An enrichment score (for each functional classification) is estimated to test for over-representation (i.e. enrichment) of target gene cohort in comparison to a reference gene cohort. It is given p-value from calculating the geometric mean of the EASE values. An annotation cluster enrichment score of 1.3 is equivalent to p-value = 0.05 and therefore annotation clusters with more than $\geq$ 1.3 are mostly reported, unless stated otherwise. Benjamini and Hochberg corrections are reported which corrects p-values to correct for multiple testing. I have not excluded cluster GO terms that were as it the statistic is conservative and enrichment scores are my main statistic of interest in identifying relevant genes that may be involved in ALS.

Chapter 6 includes gene-expression cluster-analyses using Illumina's GenomeStudio. Clusters are based on similar gene-expression signals which are integrated with the cluster analysis dendrogram with a heat-map of fold-change stratified by pathological severity. This is to identify dose-dependent changes as a consequence of pathological severity. The upper limit fold-change (2 or greater) is coloured green with the lower limit fold-change (0) coloured red. The blue arrows denote interesting clusters that show extreme changes in fold-change as we move through each severity category. Genes that show extreme alteration in fold-change are shown in the tables adjacent to the heat-map.

Gene-expression cluster analyses and dose-dependent examination of expression stratified by anatomy (Chapter 5) and stratified by spread (Chapter 6) are not reported as they were unable highlight convincing changes in expressions of ALS spread due to a lack of significant differentially expressing genes.

6.2.7 Gene expression across severity

Bioconductor R Package Lumi{Du, 2008 #1458} was utilised for additional quality control steps, which included variance stabilising transformation which makes the expression data more symmetric, and quantile normalisation. Density, cumulative distribution function, and sample hierarchal clustering plots were drawn to highlight possible abnormalities in expression data.

An additional quality control step was performed, where probe differential expression between tissue types was found to be statistically significant and therefore subsequently removed from the analysis. This was performed using Limma{Smyth, 2005 #1467}.

To test whether gene expression change was dose-dependent across different levels of severity I used a genome-wide one way ANOVA with expression level as the dependent variable and four levels of one factor (severity), which were four severity levels ascribed in the histological analyses section. Using genes with small statistical p-values from the ANOVA anlayis, I used Bioconductor package Heatmap.2 and hclust which we were able to cluster and display genes in terms of their expression profile throughout progressing severity. MFuzz, a time-based soft-clustering algorithm for gene expression, was also used to identify clusters of gene who share similar changes in gene expression as a consequence of tissue severity.

6.2.8 Bioinformatic interaction analyses

Genes found to show significant differences in the expression analysis and those highly enriched in the functional annotation cluster analysis were examined using the protein-protein network tool STRING 9.05[259] and the gene-gene and gene-protein network tool GeneMANIA[260]. Networks showing some relevance with previous ALS research are reported as I am interested in their candidacy. Follow-up analyses exploring the association and interactions between genes were performed using BioGraph[261].

6.2.9 Histological analyses

Consultant neuropathologist Dr Andrew King of the MRC London Neurodegenerative Diseases Brain Bank based at the Institute of Psychiatry, King's College London, kindly performed histological analyses at each spinal segment and the medulla for all cases and controls.

Haematoxylin and eosin (H&E) staining was used to examine the extent of loss of motor neurons in each sample. The extent of loss was categorised as follows (left column) and I collapsed these into four categories (right column) to increase statistical power and to distribute the sample size of each category equally (Figure 6-1).

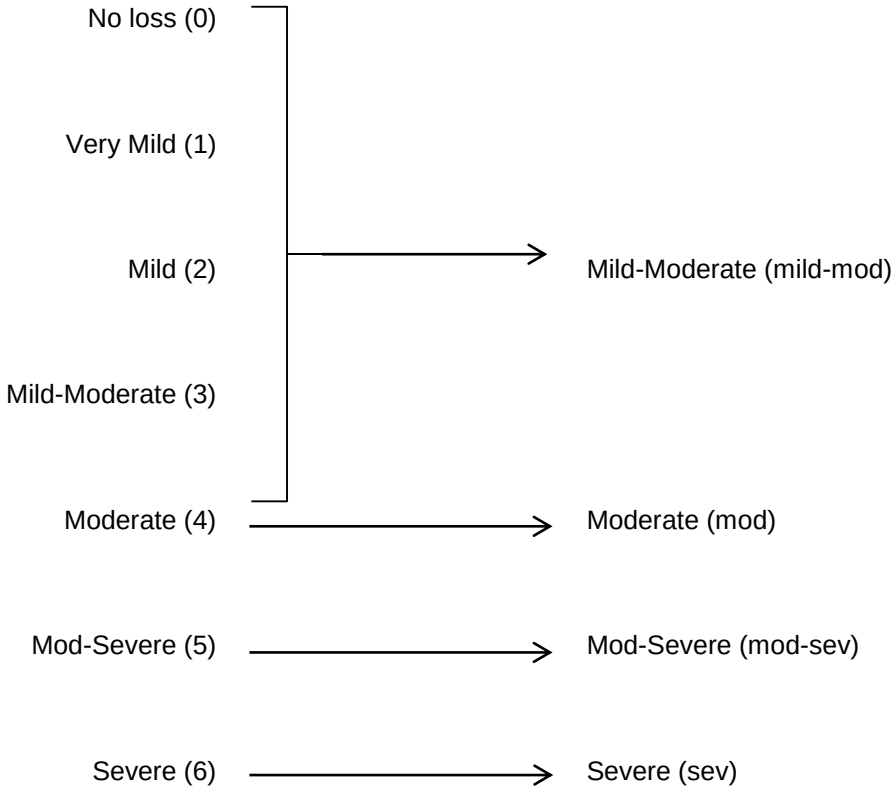


Figure 6-1. Left-hand column showing original classification of motor neuron loss severity with right-hand column showing collapsed categories used in analyses.

The number of spinal cord segments available for gene expression analysis by each category was as follows:

Mild-Moderate = 5

Moderate = 5

Mod-Severe = 7

Severe = 4

6.2.10 Clinical progression

Each case sample used in this chapter was a patient who had attended King's College Hospital. With ethical approval I examined ALS symptom progression over time to ascertain where disease spread may have occurred neuroanatomically. This entailed examining clinic notes and patient-reported details. I compared this with the histological analyses in 6.2.8.

6.2.11 Differential expression analyses

6.2.11.1 Analysis by spread

The first analysis examined differential expression and disease spread in four upper limb onset cases and three controls. I assessed genes which differentially express between cases and controls at each stage of the disease as it progresses from onset, 2nd region, 3rd region and final region (see Figure 6-2). I also aimed to identify dose-dependent effects of disease progression on gene expression for any gene identified through gene-expression cluster analysis.

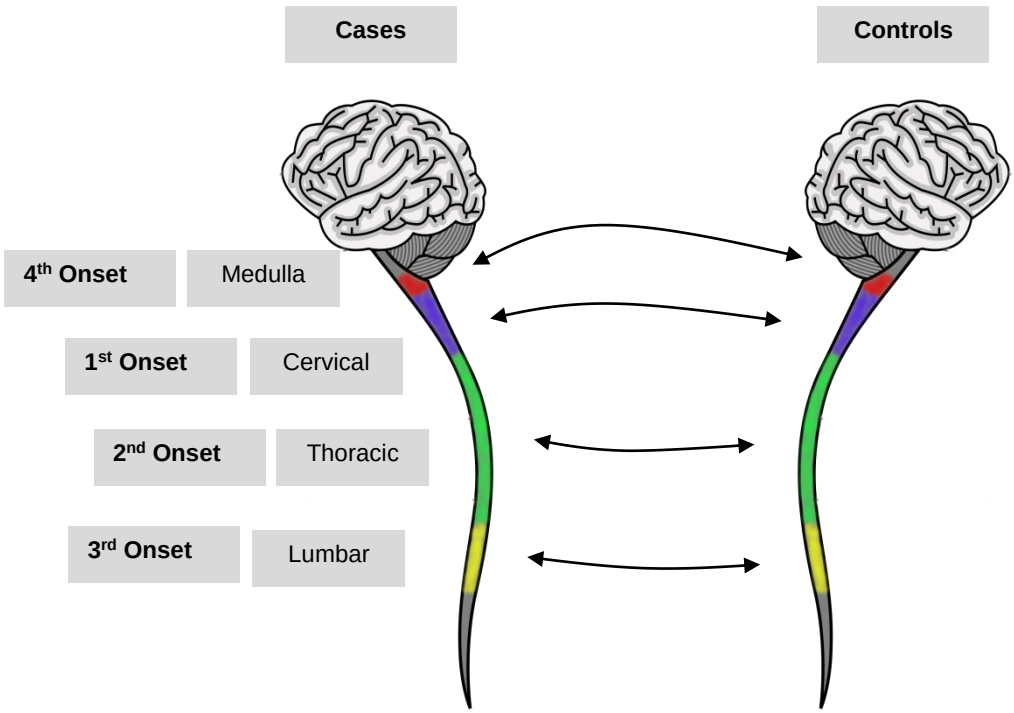


Figure 6-2. Anatomical diagram of analyses of disease spread

6.2.11.2 Analysis by pathological severity

For the second analysis I assessed genes that differentially express as a consequence of pathological severity (Figure 6-3). Pre-analysis I identified genes that naturally express differentially between spinal segments using all controls, so they could be excluded. Then I took each spinal segment and medulla for all cases, and grouped them by the four categories of pathological severity. Then I compared gene expression in cases with expression in all controls regardless of anatomical location and based only on pathological grade, excluding the genes that naturally differentially express between regions.

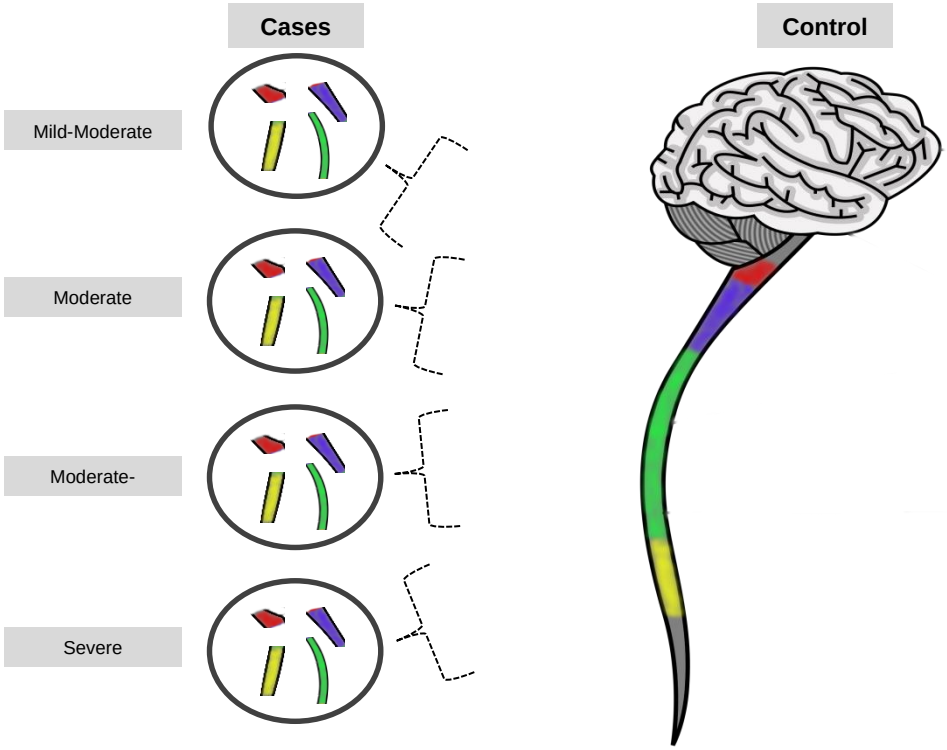


Figure 6-3. Anatomical diagram of analyses of pathological severity

6.3 Results

6.3.1 Spinal cord disease spread and severity analysis

I used clinical data and H&E staining to track disease spread in all six cases. All cases had limb onset, four with upper limb and two with lower limb. Based on previously published observations of disease spread, I predicted that for upper limb cases the disease would begin in the cervical region and spread caudally faster than rostrally[155]. Symptoms corresponded to the predicted spread pattern. Table 6-1 displays combined data for each case.

Patients with upper limb involvement (cases 2, and 4-6) followed the predicted cervical to rostral regions faster than cervical to caudal regions. For lower limb onset patients, the disease progressed in a rostral direction towards the medulla. The spread of disease through the spinal

cord followed exactly what we had predicted using previous research. Although several cases showed severe pathology that was widespread across their spinal cord and it was therefore difficult to ascertain their sequence of disease spread. This is likely to be a consequence of long disease duration at multiple spinal segments.

To demonstrate this pathological severity was rescored according to the original coding, from 0-6 for each spinal segment for the upper limb onset patients; 0 = No loss, 1 = Very Mild, 2 = Mild, 3 = Mild-Moderate, 4 = Moderate, 5 = Moderate-Severe, 6 = Severe. I excluded case ALS_2 from this analysis as they lacked both pathological and clinical data for the lumbar region which skewed the data and had unusual disease duration of 14.5 years (Figure 6-4, see appendix Figure X for equivalent figure with ALS_2 included).

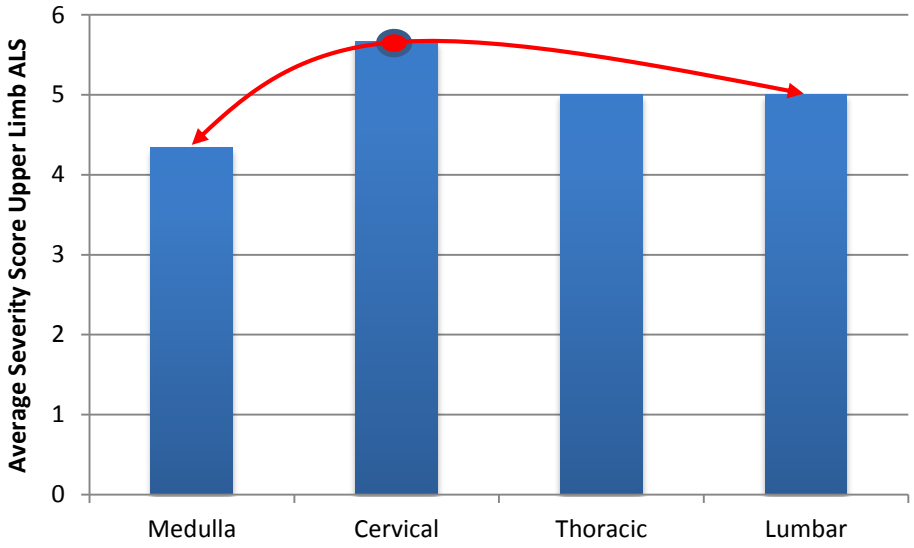


Figure 6-4. The progression of upper-limb onset ALS (n = 3) through the spinal cord using average severity scores, represented by blue bars.

The predicted pathological spread of upper-limb onset ALS is defined by the red trend-line and onset with a red-dot.

	Medulla			Cervical			Thoracic			Lumbar		
Patient	**Pathology**	**Clinical**	**Region**	**Pathology**	**Clinical**	**Region**	**Pathology**	**Clinical**	**Region**	**Pathology**	**Clinical**	**Region**
ALS_1	N/A	Final region	4th	Mod-sev	3rd region	3rd	Mod-sev	2nd region	2nd	Mod-sev	Onset	1st
ALS_2	Mild-mod	Final region	4th	Mild-mod	Onset	1st	Moderate	2nd region	2nd	N/A	3rd region	3rd
ALS_3	Very Mild	N/A	4th	Mild	N/A	3rd	Mild	N/A	2nd	N/A	Onset	1st
ALS_4	Mod-sev	N/A	4th	Severe	Onset	1st	Severe	N/A	2nd	Severe	N/A	3rd
ALS_5	Moderate	Final region	4th	Severe	Onset	1st	Mod-sev	N/A	2nd	Mod-sev	N/A	3rd
ALS_6	Moderate	N/A	4th	Mod-sev	Onset	1st	Moderate	2nd region	2nd	Moderate	3rd region	3rd

Table 6-1. Pathological and clinical information for each case by spinal segment.

The pathology column for each spinal segment displays extent of pathology and the clinical column displays disease onset and spread sequence. Using both these columns the grey column infers which region the disease began and consequently spread to in sequence of 1st to 4th.

6.3.2 Differential expression analysis of disease spread

As I could estimate the disease progression for the four upper-limb onset ALS patients, these patients were used in the proceeding analyses. The disease was estimated to move from cervical to thoracic, to lumbar and finally to the medulla. I performed differential expression analysis comparing cases (n = 4) with controls (n = 3) for each of the spinal anatomical segments in the estimate sequence of spread (see appendix table A14).

6.3.2.1 Onset region – Cervical

The most significant gene to differentially express between the cervical region of cases and controls was solute carrier family 1 (glutamate transporter), member 7 (*SLC1A7*), with p = 6.15E-04 after multiple testing adjustment using the Benjamini and Hochberg (BH) False Discovery Rate. The fold-change was 0.20 compared to controls. In Chapter 5, this gene was the most significant comparing cases and controls expression for the thoracic region, and was in the top five most significant for cervical and lumbar segments.

The protein encoded by *SLC1A7* is a glutamate and excitatory amino acid transporter, excitatory amino acid transporter 5 (EAAT5), which is vital for reducing toxic levels of extracellular glutamate. EAAT5 highly expresses in the spinal cord, temporal and prefrontal cortices, and thalamus, more-so than any other major brain or spinal cord region (UCSC Genome Bioinformatics – Genome Browser), and the retina where EAAT5 was first functionally described.

Analysis through STRING did not show any candidate or major genes already identified as important in ALS. GeneMANIA found a co-expression relationship between *SLC1A7* and the ALS candidate Sodium Channel, Beta-3 Subunit (*SCN3B*) gene (Figure 6-5). It is a regulatory subunit of voltage-gated sodium channels[296] and has been found to differentially express in ALS previously[196].

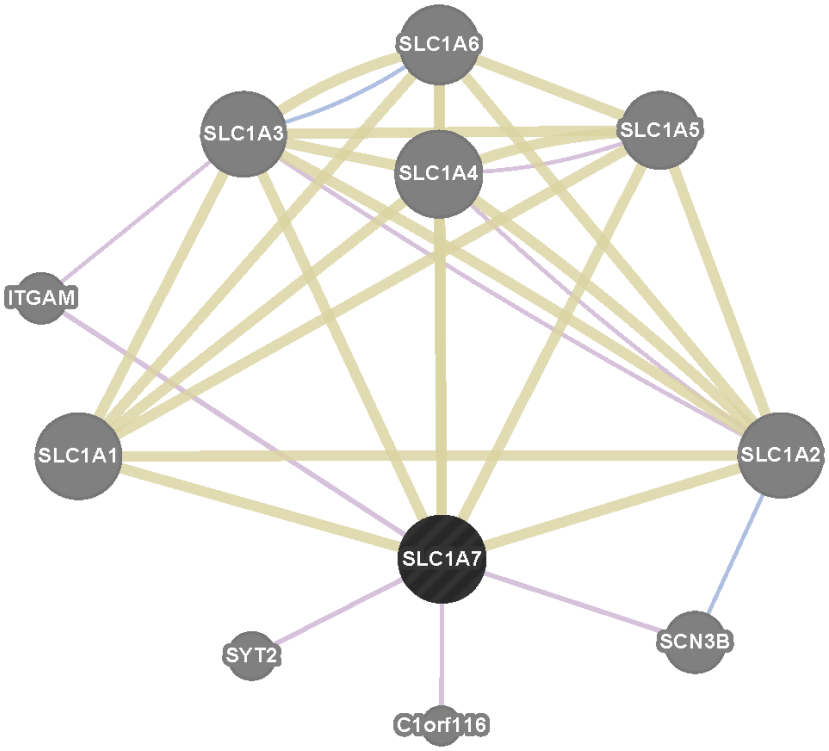


Figure 6-5. Gene-gene network of *SLC1A7* using geneMANIA. Purple line indicates a co-expression, blue line co-localisation, and a beige line shared protein domains.

Examination of *SLC1A7* dose-dependent expression as a function of disease progression did not reveal a linear step-wise reduction in fold change as the disease moved through the spinal cord (Figure 6-6). On the contrary, *SLC1A7* showed statistically significant differential expression in all spinal regions but not the medulla, although even in the medulla, the direction of change was consistent with that of the other regions.

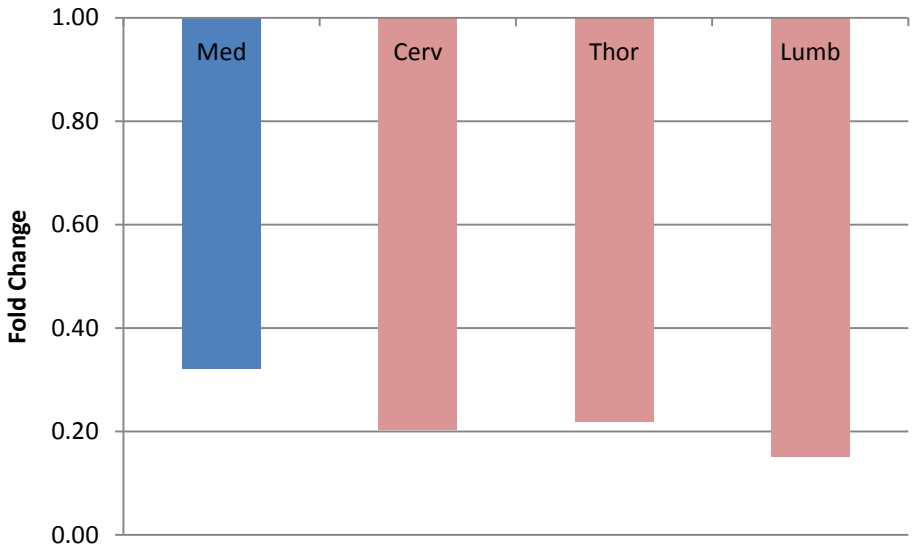


Figure 6-6. Fold-change of *SLC1A7* in each spinal cord segment.

The pink bars represent that *SLC1A7* showed significant differential expression at these segments, blue bars represents non-significance.

The second and last most significant gene was Pote Ankyrin Domain Family, Member B (*POTEB*), which has no well-characterised function. *POTEB* showed no network associations with any ALS gene and has not been implicated in ALS previously.

6.3.2.2 Second region – Thoracic

The most significant gene to differentially express was, as described for the cervical region, *SLC1A7* (p = 1.90E-06), with a fold-change of 0.22 (Figure 6-6). The second and last significant gene was Interleukin 7 Receptor (*IL7R*), which has no known association with ALS and revealed no interesting network connection. *IL7R* synthesises a glycoprotein that helps regulate lymphopoiesis. This will become important later on, when we examine pathological severity.

6.3.2.3 Third region – Lumbar

The most significant gene to differentially express between cases and controls was again *SLC1A7* (p = 2.89E-10), with a fold-change of 0.15 (Figure 6-6). Table 6-2 displays the genes of the lumbar segment that showed statistically significant differential expression when compared with controls. None of these genes have previously been implicated in ALS or show a functional network association with an ALS gene.

Gene name	Symbol	Fold Change	Adj. P-Value	Function
Solute Carrier Family 1 (glutamate transporter), Member 7	*SLC1A7*	0.151	2.89E-10	*Glutamate transporter*
Interleukin 7 Receptor	*IL7R*	3.165	3.11E-07	*Regulation of lymphopoiesis*
Quinolinate Phosphoribosyltransferase	*QPRT*	1.945	4.68E-06	*Catabolises quinolinic acid*
Acyl-CoA Synthetase Family Member 5	*ACSM5*	2.557	4.28E-05	*Involvement in energy storage and metabolism*
5', 3'-Nucleotidase, Cytosolic	*NT5C*	1.917	1.50E-04	*Catalyses dephosphorylated C4H4N2 monophosphates*
Alkaline Ceramidase 2	*ACER2*	0.445	1.14E-02	*Generates sphingosine for ITGB1 and cell adhesion*
Dihydrofolate Reductase-like 1	*DHFRL1*	1.685	1.40E-02	*Enyme for folate metabolism*
Dynein, Axonemal, Assembly Factor 2	*DNAAF2*	1.752	4.53E-02	*Helps motility in cilia and flagella*
Recoverin	*RCVRN*	0.471	4.69E-02	*Helps generate neuronal calcium sensors*

Table 6-2. Genes that significantly differentially expressed between cases and controls at the third onset region

6.3.2.4 Last region – Medulla

The gene expression profile of the Medulla is somewhat different from other spinal cord anatomical profiles (as shown in Chapter 5). Table 6-3 displays the genes found to differentially express between upper-limb onset cases and controls. None of the genes have been previously implicated in ALS or showed functional network connection with known ALS genes.

Gene Name	Symbol	Fold Change	Adj. P-Value	Function
Family with Sequence Similarity 43, Member A	*FAM43A*	0.55	2.11E-04	Unknown
MCM3AP Antisense RNA 1	*MCM3AP-AS1*	3.33	1.06E-03	Non-protein coding anti-sense RNA
Shroom Family Member 2	*SHROOM2*	0.63	9.60E-03	Amiloride-sensitive sodium channel activity
SRY (sex determining region Y)-box 17	*SOX17*	0.60	9.60E-03	Transcription factor and role in embryogenesis
Sulfotransferase Family, Cytosolic, 1B, Member 1	*SULT1B1*	0.46	1.34E-02	Catalyse sulphate conjugation of hormones, drugs
Centromere Protein V	*CENPV*	0.60	3.26E-02	Centromere organisation; distributes heterochromatin

Table 6-3. Genes that significantly differentially expressed between cases and controls at the last onset region

6.3.3 Differential expression analysis of pathological severity

The analysis in the previous section showed relatively uniform gene expression with little difference between each region as the disease progressed. For the analysis in this section I categorised each spinal region, including Medulla, by pathological severity. Creating four categories of severity, I compared each category with the all regions from all the controls. I also compared gene expression between each region, using all cases and controls, to obtain and control for genes that naturally differentially expression between the regions I was examining. I then excluded these genes from the below analyses (see appendix table A15).

6.3.3.1 Mild-moderate pathology

Differential expression analysis of regions with Mild-Moderate pathology between cases (n = 5) and controls (n = 12) revealed 333 genes that differentially expressed. Table 6-4 shows the top five of these genes (for all genes see Appendix Table A16).

Gene Name	**Symbol**	**Fold Change**	**Adj. P-Value**	***Function***
FBXL19 Antisense RNA 1 (head to head)	*FBXL19-AS1*	0.19	2.16E-33	*Possible non-coding RNA*
Family with sequence similarity 21, Member A	*FAM21A*	0.08	9.50E-19	*Endosomal activity*
Major Histocompatibility Complex, class II, DR beta 4	*HLA-DRB4*	0.09	2.56E-15	*Membrane-based antigen presenter*
MCM3AP antisense RNA 1	*MCM3AP-AS1*	3.54	8.81E-11	*Unknown*
SRY (sex determining region Y)-box 18	*SOX18*	0.41	4.16E-10	*Transcription regulation of embryogenesis*

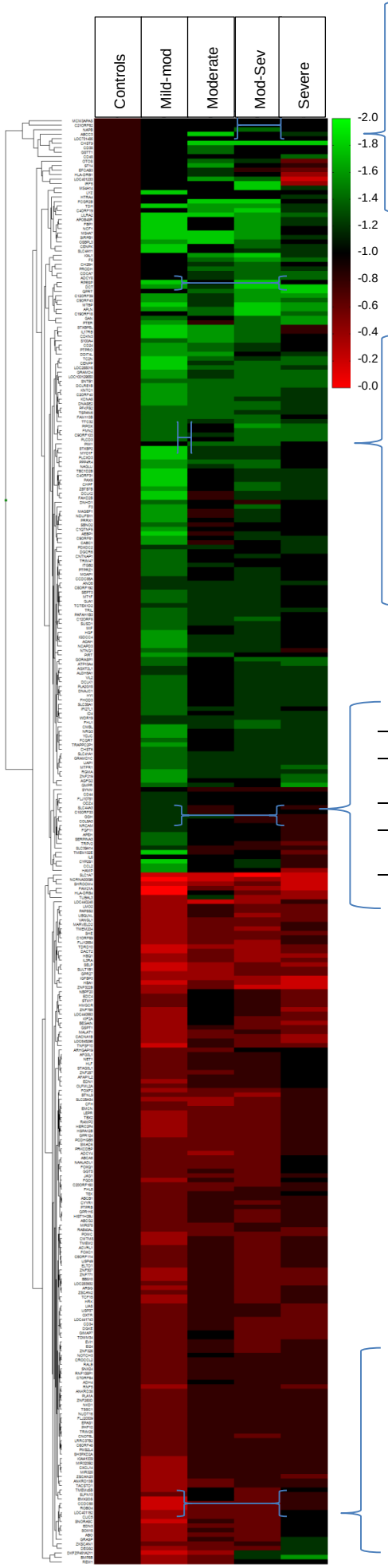
Table 6-4. Top five genes out 333 at Mild-Moderate spinal segments showing significant differential expression

Cluster-analysis of gene expression was performed for all 333 genes from this category. The dendrogram and heat-map (Figure 6-7) shows these clusters with corresponding fold-change as we move from Mild-Moderate pathology to Severe pathology. The genes in this diagram are the 333 which significantly differentially expressed in the Mild-Moderate pathology category. None of these genes have previously been implicated in ALS and none showed relevant gene-gene or gene-protein network relationships with previously established in genes.

Using DAVID functional annotation clustering for all 333 genes derived 14 clusters with an enrichment score greater than 1.3, where 12 contained terms that were significant after multiple testing corrections. See Table 6-5 for the cluster headings, and appendix A13 for the entire list of cluster significant genes.

Main Functional Clusters	***Enrichment Score***	***Highest Gene count***
Blood vessel development	3.42	14
Glycoprotein signalling	2.86	95
Glycoproteins & transmembrane activity	2.84	114
Fibronectin	2.69	11
Leukocyte migration & cell motility	2.65	15
Leukocyte migration & response to external stimuli	2.31	8
Defence response	2.27	20
Surface antigen & binding	2.06	14
Extracellular region	2	46
Plasma membrane	1.71	47
Leukocyte migration & response to bacterial molecule	1.67	7
Phosphate activity	1.66	7
Leukocyte migration & chemotaxis	1.51	12
EGF-like domain	1.41	12

Table 6-5. Functional annotation clustering with enrichment score > 1.3 of 333 genes found to differentially express at regions with Mild-Moderate pathology.



Gene	Function
OTOS	Neurosensory epithelium of the inner ear
ST14	Degrades extracellular matrix
EFCAB3	Calcium-binding
HLA-DRB1	Contributes to antigen presenting molecules
LOC401233	Pseudo gene
IRF5	Transcription factor for interferons IFNA and INFB

Gene	Function
STXBP5L	Likely vesicle trafficking and exocytosis
IL17RB	Proinflammatory cytokine receptor for IL17
CDKN3	Possible role in cell cycle regulation

Gene	Function
TBC1D2B	GTPase-activating protein
C4ORF31	Promotes neuron migration, growth & survival
PAX6	Master control gene for sensory organs
CHPF	Transferase activity
ZBTB7B	Transcription regulation of immature T-cell precursors
DCLK2	Possible down-regulation of CRE-dependent activation
FAHD2B	Possible role in hydrolase activity

Gene	Function
TMEM132E	Transmembrane protein, function undetermined
IL6	Important for fever and of the acute phase response.
CYP2S1	Possible extrahepatic xenobiotic metabolism
CCL2	Recruitment of white blood & T-Cells for inflammation
HAMP	Signalling molecule for iron homeostasis

Figure 6-7. Gene-expression cluster analysis and heat-map.

Examining fold-change across pathological severity categories for genes that differentially expressed in the Mild-Moderate category. X-axis (top) is pathological severity category and y-axis (left) is gene.

Gene	Function
ZKSCAN1	Transcription regulation
DEGS2	Sphingolipid enzyme
DKFZP451A211	Uncharacterised
BMP8B	Functions in cartilage and bone formation
REM1	Endothelial cell sprouting & cytoskeletal reorganization

One of the genes found to differentially express, Sushi domain containing 1 (*SUSD1*), has been previously implicated in ALS in a genome-wide association study in 2007[245]. The finding has not yet been replicated. The sequence and putative function of *SUSD1* is similar to other proteins containing EGF-like domains, involved in calcium-binding and immune response and apoptosis. In this thesis the *SUSD1* gene belonged to the EGF-like domain cluster.

Using GeneMANIA and STRING there were no significant interactions between any of the 333 genes here and known major ALS genes. Non-replicated ALS genes *CRIM1*, *KDR*, and *RAMP3* showed network relationships with genes involved in blood vessel development and glycoprotein function, mostly weak interactions through co-expression, co-localisation, and shared protein domains. Most of the interactions were indirect. It is reasonable to expect some low-level interaction between ALS genes (n = 110) and the genes found here (n = 333), as 333 genes is nearly 2% of all human genes.

6.3.3.2 Moderate pathology

Analysis of pathologically Moderate spinal regions (n = 5) revealed 34 genes that differentially expressed. All significant genes are display in the gene expression cluster-analysis heat-map (Figure 6-9) and the top five are displayed in Table 6-6 (for all genes see Appendix Table A17).

GEMIN5 physically interacts with putative ALS gene Survival Motor Neuron 1 (*SMN1*) to create protein complexes which are important for the synthesis of cytoplasmic small ribonucleoproteins (snRNPs) and the splicing of nucleic pre-mRNA (Figure 6-8).

The gene-expression cluster-analysis revealed four genes that showed Moderate dose-dependent alteration in fold-change between pathological categories. All four genes are involved in immune responses. No gene had been implicated or had a network interaction with a previous ALS gene.

In the functional annotation analysis the top cluster had an enrichment value of 1.2 and was again involved in glycosylation (see Appendix Table A17), although no term was significant after correcting for multiple testing. I found no other notable interaction with genes differentially expressing at regions showing Moderate pathology.

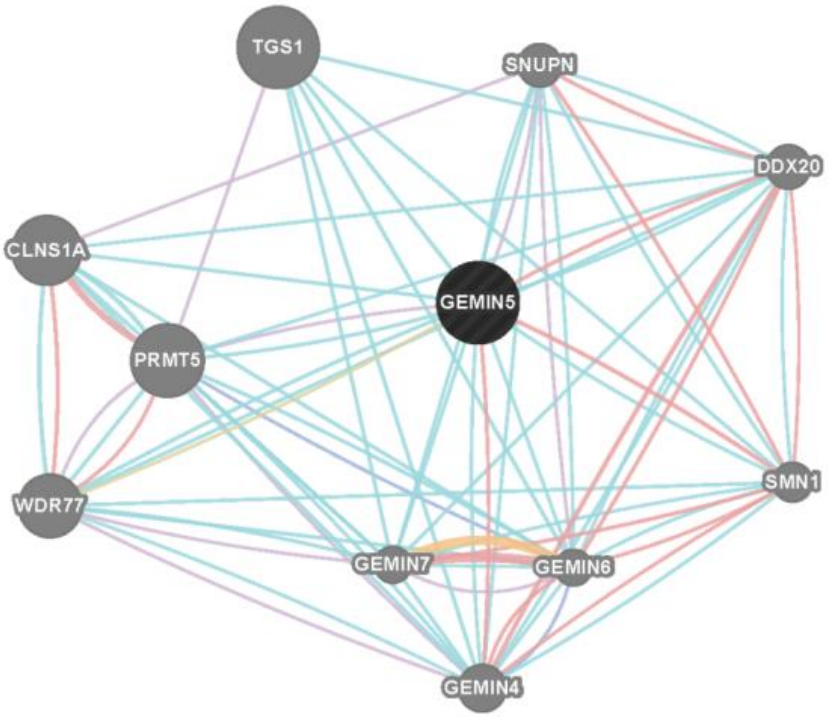


Figure 6-8. Gene-gene network of *GEMIN5* using geneMANIA.

Purple line indicates a co-expression, pink line physical interactions, orange line predicted interaction, light blue line shared pathway, dark blue line co-localisation, and a beige line shared protein domains.

Gene Name	Symbol	Fold Change	Adj. P-Value	Function
Solute Carrier Family 1 (glutamate transporter), Member 7	*SLC1A7*	0.26	2.96E-05	*Glutamate transporter*
Leucine-Rich repeat LGI Family, Member 4	*LGI4*	0.58	1.04E-04	*Encodes for leucine-rich protein in brain*
Major Histocompatibility Complex, Class I, A	*HLA-A29.1*	0.17	3.36E-04	*Antigen presenter*
Gem (nuclear organelle) Associated Protein 5	*GEMIN5*	0.62	2.05E-03	*Forms cytoplasmic complexes with SMN proteins*
Carboxypeptidase A4	*CPA4*	0.45	2.53E-03	*Helps release carboxy-terminal amino acids*

Table 6-6. Top five genes out 34 at Moderate spinal segments showing significant differential expression

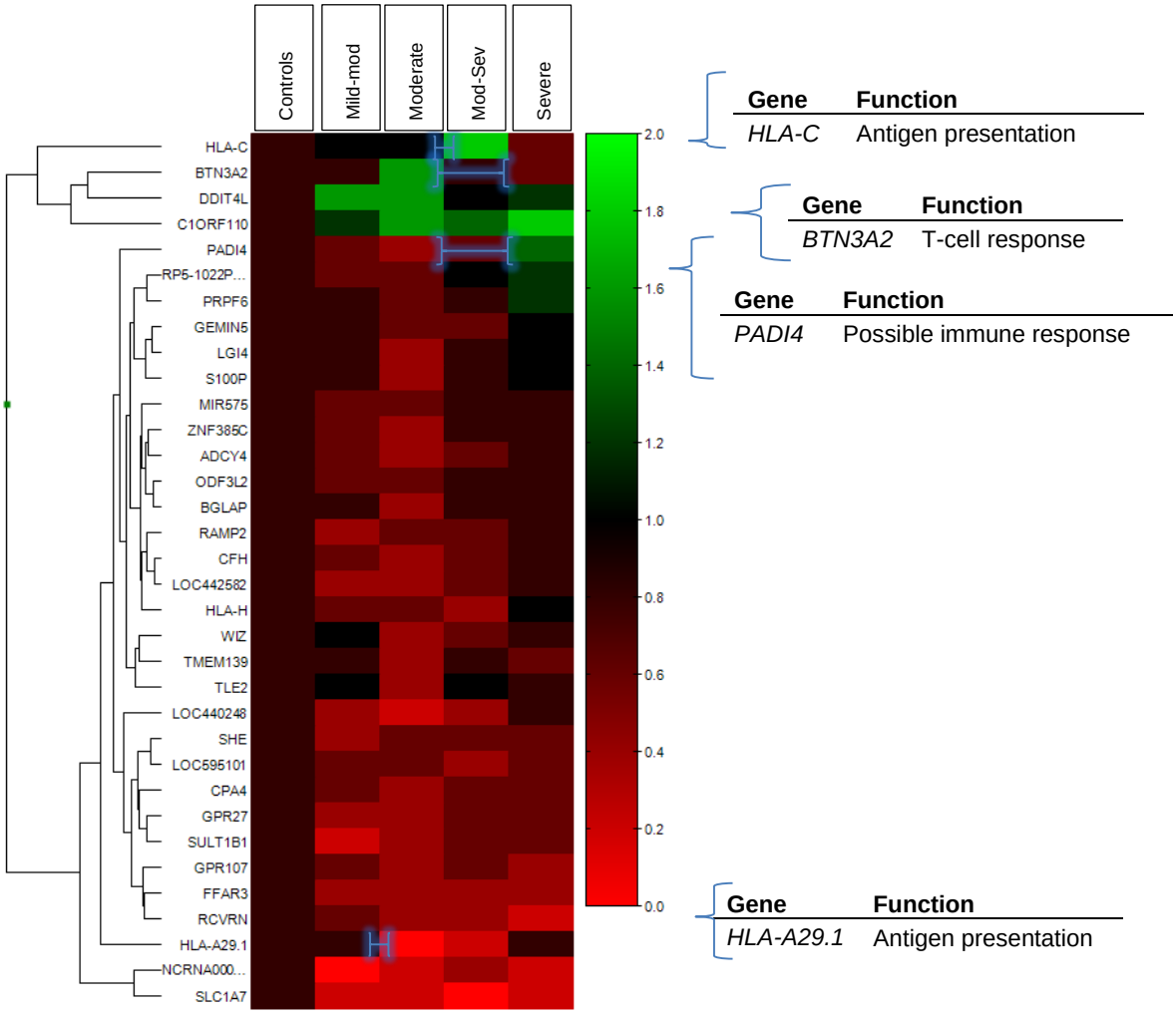


Figure 6-9. Cluster analysis and heat-map of significant differentially expressing genes. Examining fold-change in the cases with Moderate pathology. X-axis is pathological severity category and y-axis is gene.

6.3.3.3 Moderate-severe pathology

Analysis of pathologically Moderate-Severe spinal regions (n = 7) revealed 138 genes that differentially expressed. All significant genes are displayed in the gene expression cluster-analysis heat-map (Figure 6-10) and the top five are displayed in Table 6-7 (for all genes see Appendix Table AX).

Gene Name	Symbol	Fold Change	Adj. P-Value	Function
Solute Carrier Family 1 (glutamate transporter), Member 7	*SLC1A7*	0.15	2.16E-33	*Glutamate transporter*
Olfactory Receptor, Family 7, Subfamily E, Member 125 pseudogene	*PJCG6*	0.44	2.85E-06	*Pseudogene mRNA*
Mdm2, Transformed 3T3 cell double minute 2, p53 Binding Protein, 104kDa	*MTBP*	1.92	7.74E-06	*Cell cycle regulation and tumour suppression*
Hyaluronan and Proteoglycan Link Protein 3	*HAPLN3*	0.46	6.73E-05	*Hyaluronic acid binding*
Adenosine Deaminase	*ADA*	0.50	7.24E-04	*Hydrolyses of adenosine to inosine*

Table 6-7. Top five genes out 34 at Moderate-Severe spinal segments showing significant differential expression

The gene-expression cluster-analysis revealed very few genes that showed dose-dependent changes in gene expression (see Figure 6-10). The only gene which showed some decrease in expression as pathology worsened was Rhox Homeobox Family, Member 1 (*RHOXF1*). This gene is not implicated in ALS and shows no network association with ALS genes.

The functional annotation analysis top cluster was involved in metabolic and biosynthetic processes, with the latter having an enrichment value of 1.39 (see Appendix Table A18). No cluster was significant after correcting for multiple testing. *MSRA*, which was grouped in the metabolic processes category, showed an indirect association with ALS which has been discussed in Chapter 5.

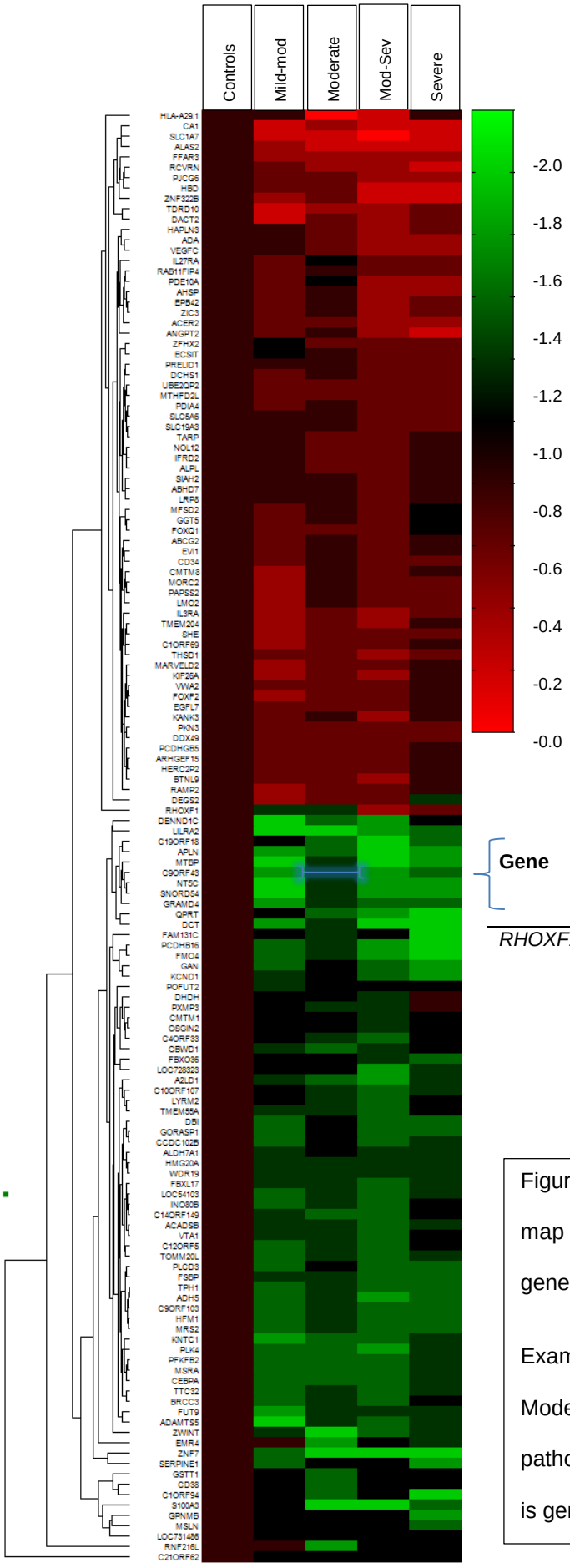


Figure 6-10. Cluster analysis and heat-map of significant differentially expressing genes.

Examining fold-change in the cases with Moderate-Severe pathology. X-axis is pathological severity category and y-axis is gene.

6.3.3.4 Severe pathology

Analysis of pathologically Severe spinal regions (n = 4) revealed 205 genes that differentially expressed. All significant genes are displayed in the gene expression cluster-analysis heat-map (Figure 6-12) and the top five are displayed in Table 6-8 (for all genes see Appendix Table A19).

Gene Name	Symbol	Fold Change	Adj. P-Value	Function
Angiopoietin 2	*ANGPT2*	0.36	1.55E-13	*Disrupts vascular remodelling & induces apoptosis*
epididymal sperm binding protein 1	*ELSPBP1*	3.44	2.43E-08	*Sperm-coating protein*
Inositol Polyphosphate-4-Phosphatase, Type II, 105kDa	*INPP4B*	1.68	2.50E-07	*Phosphatidylinositol signalling pathway*
Olfactory Receptor, Family 7, Subfamily E, Member 125 Pseudogene	*OR7E125P*	0.43	4.87E-07	*Pseudogene mRNA*
Transketolase-like 1	*TKTL1*	2.03	1.53E-06	*Links pentose phosphate & glycolytic pathways*

Table 6-8. Top five genes out 34 at Severe spinal segments showing significant differential expression

The Angiopoietin 2 (*ANGPT2*) protein closely interacts with the protein of ALS gene *VEGFA*. The relationship is believed to be context dependent with two relatively opposing functions; (a) through direct binding, and/or in the absence of *VEGFA*, it is believed to inhibit expression of *VEGFA* inducing endothelial cell apoptosis. On the other hand (b) the relationship facilitates endothelial cell proliferation through *VEGFA* activation of *ANGPT2* (Figure 6-11).

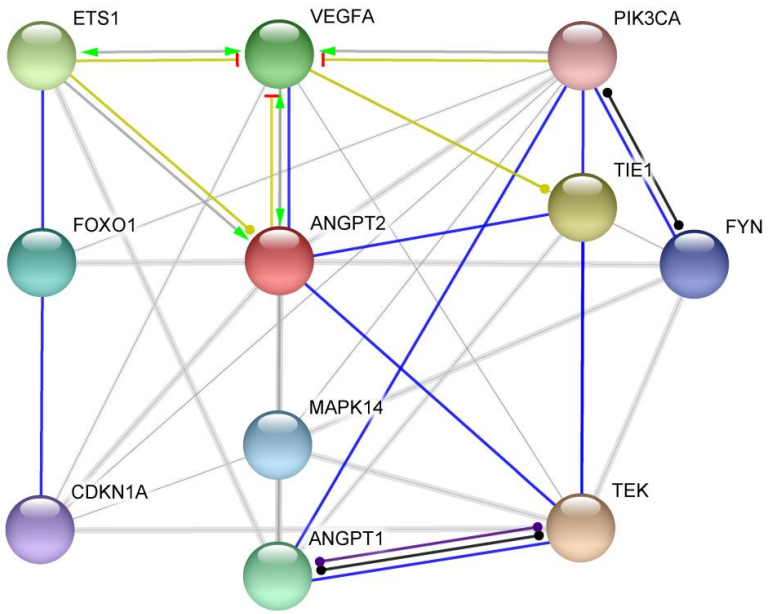


Figure 6-11. Protein network of *ANGPT2* using STRING.

Green arrows represent activation, blue lines binding, yellow co-expression, red line-end inhibition, and grey lines putative interaction through searching experiments, databases, text-mining, and homology.

The gene-expression cluster analysis revealed 18 genes with altered activity corresponding to categories of pathological severity. These genes are highlighted in the tables within Figure 6-13. No gene listed has previously been found to contribute to the disease, but do implicate ALS genes through network associations. The network analyses of *DNAH2* revealed gene-gene interaction in sharing a pathway with Dynactin 1 (*DCTN1*) (Figure 6-12). *DCTN1* produces the multifunctional

protein Dynactin, which when dysfunctional in ALS is believed to impair axonal transport supporting the axonopathy hypothesis of pathogenesis[297] (Figure 6-12).

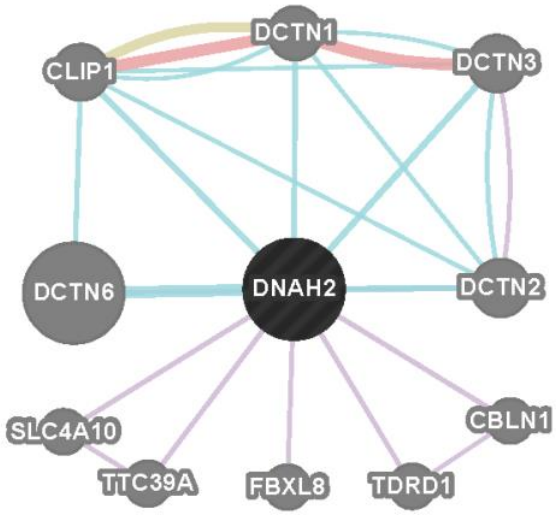
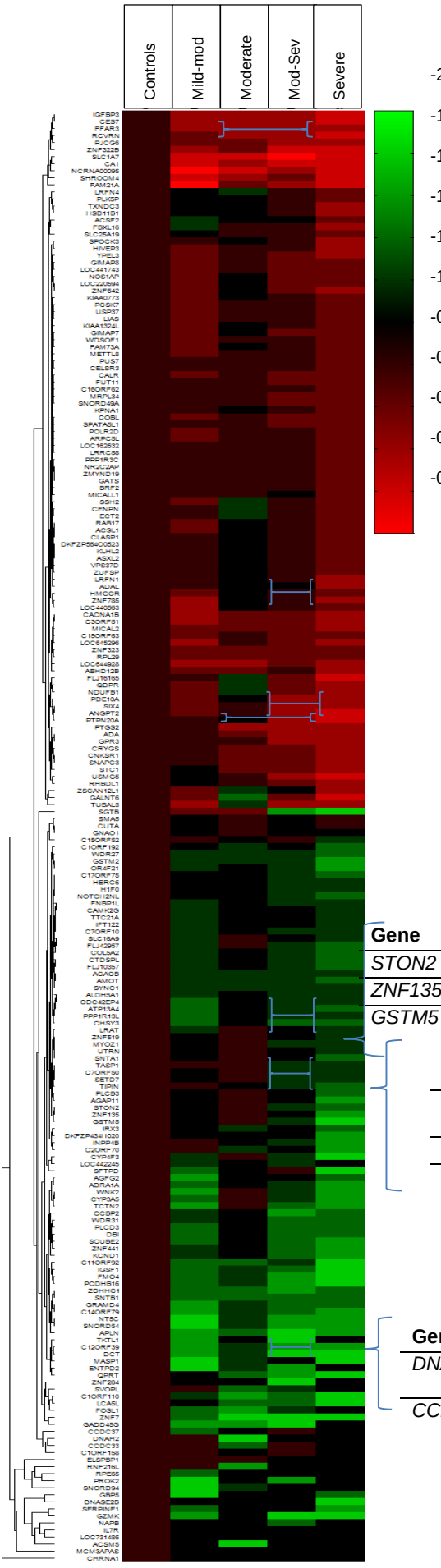


Figure 6-12. Gene-gene network of *DNAH2* using geneMANIA.

Purple line indicates a co-expression, pink line physical interactions, light blue line shared pathway, and a beige line shared protein domains.

No functional annotation cluster had an enrichment score greater than 1.3. The highest two clusters had a score of 1.17 and the third 1.12. Cluster terms were striated muscle cell development, microsomes and vesicular fraction, and fatty acid metabolic processes (see appendix table A19). There were no other meaningful interactions with the genes differentially expressing at regions showing Severe pathology and known ALS genes.



Gene	Function
ACSF2	Catalyst in fatty acid metabolism
FBXL16	Ubiquitination of cell-cycle proteins

Figure 6-13. Cluster analysis and heat-map of significant differentially expressing genes.

Examining fold-change in the cases with Moderate-Severe pathology. X-axis is pathological severity category and y-axis is gene.

Gene	Function
FLJ16165	Unknown
QDPR	Recycling enzyme for BH4
NDUFB1	Subunit of NADH dehydrogenase

Gene	Function
ZSCAN12L1	Pseudogene
GALNT6	Oncofetal fibronectin synthesis
TUBAL3	Major component of microtubules

Gene	Function
SGTB	Regulates HSC70 & HSP70 ATPase activity

Gene	Function
STON2	Possible vesicle recycling
ZNF135	Cell morphology & cytoskeletal regulation
GSTM5	Conjugation of glutathione to hydrophobic electrophiles

Gene	Function
CYP4F3	Contributes to a leukocyte chemoattractant
LOC442245	Pseudogene
SFTPD	An innate immune system collectin

Gene	Function
DNAH2	Respiratory cilia protein involved in sperm motility
CCDC33	Testis-expressed possible peroxisomal protein.

6.3.4 Changes in gene expression as a consequence of severity

Using a genome-wide ANOVA of gene expression by severity, there four genes significant after BH multiple testing correction. These were *ST5*, *PARD3B*, *CNTFR*, and *RNF216L*. None of these have been previously implicated in ALS.

63 genes with a non-adjusted p-value < 0.001 (see Appendix Table A20) were selected, and I performed cluster analysis represented by a heatmap (Figure 6-14). This included two genes that have been previously implicated in ALS, DIAPH3{Daoud, 2011 #1378} via resequencing and VEGFA via candidate gene approach{Oosthuyse, 2001 #382}.

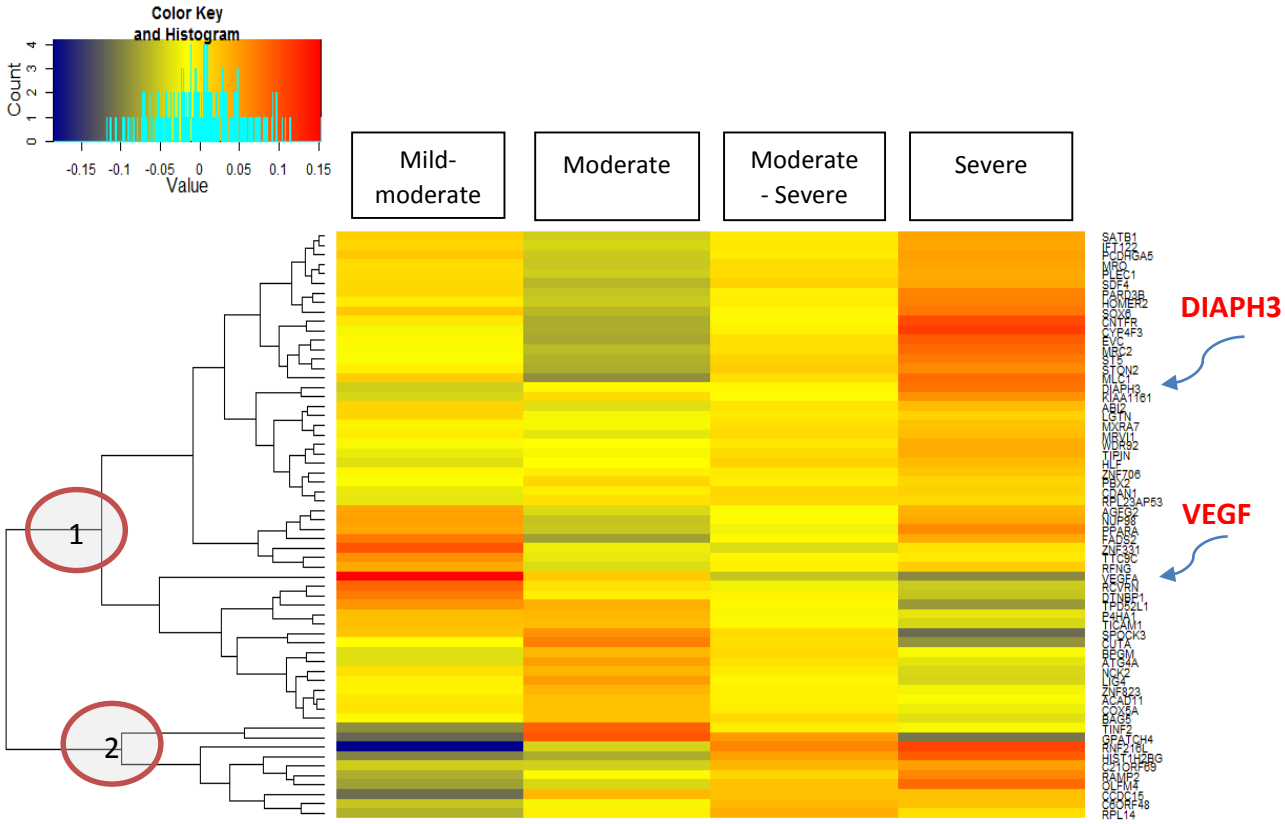


Figure 6-14. Cluster analysis and heat-map of genes that show dose-dependent gene expression with severity level.

Genes in clusters circled 1 and 2 were analysed using DAVID's functional annotation analyses. Cluster 1 showed a gradual increase in intracellular protein transport and localisation, with an involvement of 37 genes and a resulting enrichment score of 1.1. Cluster 2 showed an up-down

change in leukocyte response, with an involvement of 12 genes and a resulting enrichment score of 1.22.

I also performed soft-clustering analyses on patterns of gene expression change through each severity (Figure 6-15).

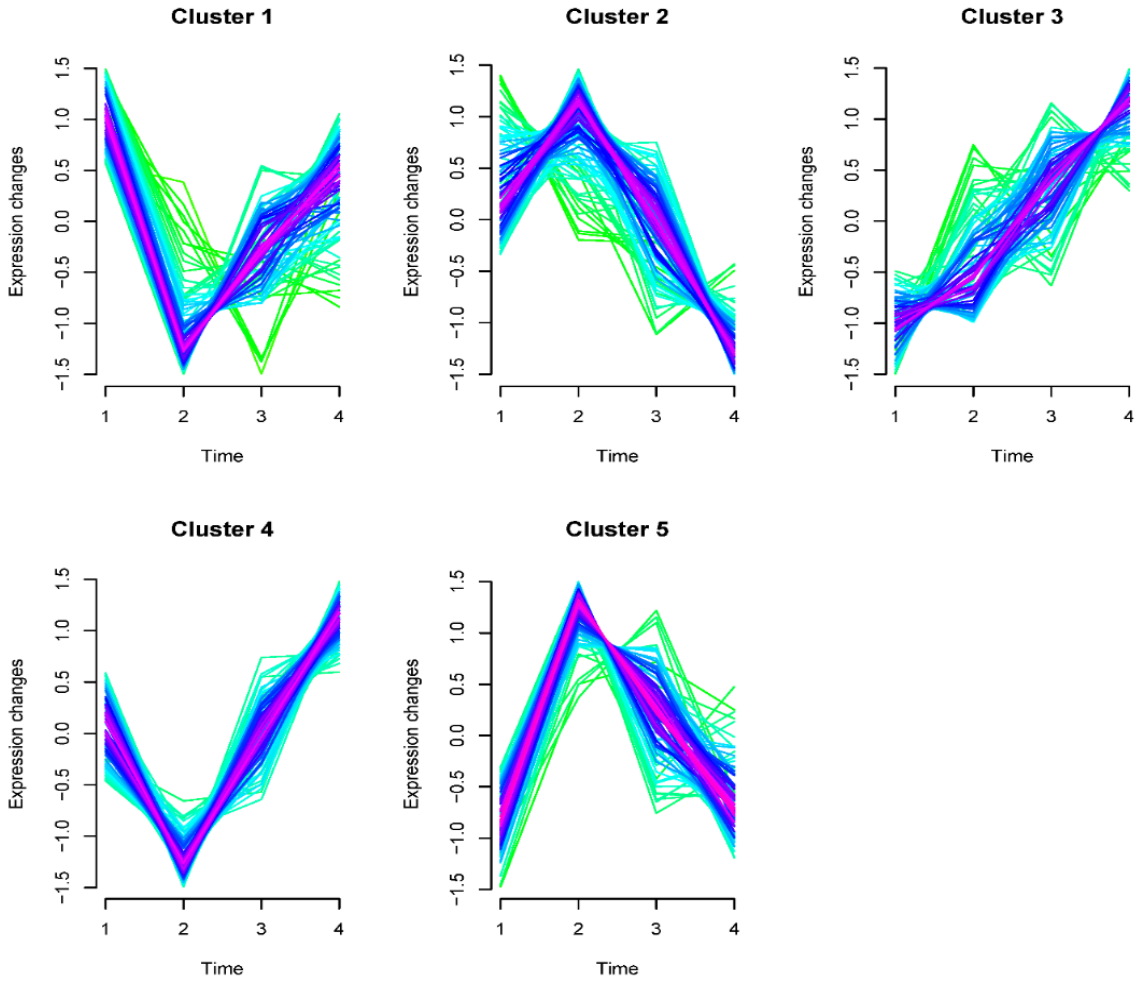


Figure 6-15. Cluster analysis of changes in expression through severity. Time can be supplemented by severity level (where 1 = mild-moderate and 4 = severe). The Y-axis represent expression changes. Colour of line represent distance from cluster centre.

Each cluster was examined using DAVID's functional annotation analyses. Cluster 1 was involved in regulation of GTPase activity (enrichment score = 1.64), Cluster 2 in Regulation of leukocytes

proliferation (enrichment score = 1.52) and regulation of transcription activity (enrichment score = 1.39), Cluster 3 in nucleic methylation (1.6), Cluster 4 in calcium related cell-adhesion (1.67) and Cluster 5 in nucleic lumen function (1.56).

6.4 Discussion

Through the analysis of gene expression and disease spread, inferred via clinical and pathological information, I have identified significant differences in expression between cases and controls.

Analyses of gene expression by spread found only two genes to differentially express for cervical (onset region) and thoracic (second region) regions, four genes for the lumbar (third) region, and five genes for the medulla (final) region.

Analyses of gene expression by pathological severity also identified significant differences between cases and controls. The number of differentially expressing genes ranged between 34 and 333 according to pathology category, and each category revealed an expression profile analysed by functional annotation enrichment to vary from all others. The functional categories of pathologically Mild-to-Moderate samples were multifarious with 14 gene clusters highly enriched, covering functions such as blood vessel development, glycoprotein signalling and leukocyte activities. As the severity increased there were fewer annotated cluster. These clusters showed a reduced enrichment score but were associated with glycosylation for Moderately pathological samples, metabolic processes for Moderately-Severe samples, and striated muscle development, microsome involvement, and vesicular fractions in the Severely pathological samples.

One gene showed expression changes in all three spinal regions excluding medulla, and all four categories of severity: *SLC1A7*. Evidence shows that this gene is less expressed in the Medulla in comparison to the other spinal regions examined in this thesis. *SLC1A7* is discussed in the next section.

6.4.1 The spread of ALS

Spread of ALS in upper-limb onset cases was estimated to have onset in the cervical regions and moved caudally faster than it did rostrally. However, this was not reflected in the gene expression

analyses. The gene expression was relatively uniform throughout the spinal cord, with the exception of *SLC1A7*, which was under-expressed very significantly in cervical, thoracic and lumbar regions.

SLC1A7 has been implicated in ALS previously in a comparative genomic hybridisation study examining genome-wide copy number variation in 71 sporadic ALS cases and 700 non-ALS controls[298]. The most promising gene was SLC1A7, which showed a copy number gain found exclusively in one patient[298]. In a recent study SLC1A7 was up-regulated with the authors reasserting its candidacy due its likely functional involvement in ALS pathology[299].

SLC1A7 protects motor neurons from excitotoxicity by transporting extracellular glutamate[300] through synthesising the protein EAAT5. EAAT5 shares 36% sequence identify with glia protein *EAAT2*, which is encoded by *SLC1A2* of the same gene family and has almost the exact same function as EAAT5. EAAT5 is considered to be a retinal specific glutamate transporter[301] but it also highly expresses in the spinal cord, temporal and prefrontal cortices, and thalamus, more-so than any other major brain or spinal cord region (UCSC Genome Bioinformatics – Genome Browser).

EAAT2 expression levels have been found to be significantly reduced in brain and spinal cord in approximately 60% of ALS patients[302]. The cause of *EAAT2* reduction may be alternative aberrant splicing[303] but this has not yet been confirmed. An alternative hypothesis is that it is indirectly involved in ALS. Knockdown of TDP-43 in drosophila, which causes an ALS cytological phenotype, significantly affects glia EAAT1 and *EAAT2* expression[304]. It is plausible that *EAAT2* expression either delays or exacerbates ALS onset, but does not cause ALS, as has been seen in mice[305].

EAAT2 dysfunction, or perhaps inefficiency, most likely contributes to ALS via glia. It is also possible that EAAT5 is expressing from glia in our data as there was no dose-dependent change across different stages of onset. We would expect almost non-existent expression in regions

where onset had occurred early, because by the time there is severe involvement, motor neurons would be very few in number, whereas glia counts would be almost unchanged. In the work presented here, the cervical and thoracic regions were where ALS manifested for the longest period of time. Therefore, in these regions, we would expect near non-existent expression of known ALS genes in these regions in which we did not, nor did we find any other gene showing differential expression.

The expression profile of the Medulla was different from the other spinal regions. The number of genes found to differentially expression was small compared to previous research with ALS spinal regions. This may be a power issue but a more likely explanation is that that the disease duration for the cases used in this analysis was above average. This meant that the disease had manifested for a long period of time in the second, third and fourth regions, which could make the pathology of this region indistinguishable from the first onset region. The H&E pathological evidence does show some decrease in pathological severity, but this is only slight. An alternative explanation is that there is no changing gene expression as the disease progresses, but this would not be concordant with previous gene expression work in ALS and other neurodegenerative diseases.

The medulla presented a different gene expression profile from the spinal cord in relation to ALS. It could be that the disease manifests itself differently here and the background expression profile is naturally different between the medulla and the spinal regions examined. When comparing gene expression between spinal regions in controls, the Medulla shows many significantly differentially expressing genes when compared with non-Medulla regions. It is anatomically very different than the spinal regions examined in this thesis and it may be very different functionally for ALS pathology. Its involvement in ALS is likely due to its proximity of ALS initial onset and rostral spread.

6.4.2 Pathological severity and ALS

The changes in gene-expression profiles as the disease travels from its early beginnings to later stages should be informative about its pathogenesis. I would expect that at more mildly affected regions gene expression profiles reflect motor neural responses to the disease, whereas more severe pathologically effected regions reflect glia-based responses.

6.4.2.1 Mild-moderate pathology

The least severe group, with mild to moderate pathology defined by loss of motor neurons, had the most genes showing differential expression. It showed high enrichment for 14 annotation clusters (see Table 6-5). These cover a vast range of functions but present us with a picture of gene activity just after the spinal cord begins to show pathology.

Blood vessel development was the most enriched cluster. This cluster also contained angiogenesis as an important function, with nine of the 14 genes loading onto this category. The association of two functions with ALS is not new. There is strong evidence implicating the Angiogenin gene (*ANG*) as important in ALS, with mutations causing sub-type ALS9[41]. However, blood vessel development and angiogenesis were first implicated through the identification of the vascular endothelial growth factor A (*VEGFA*) gene [106] as associated with ALS. This gene codes for a growth factor and signalling protein for *ANG*. In the most Severe group there was also an relationship between *VEGFA* expression and the most significantly expressed gene from the cluster, *ANGPT2*, suggesting that blood vessel development and angiogenesis was a major factor for these cases throughout the course of their disease. Figure 6-14 shows the predicted relationship between *ANGPT2* and *ANG* through angiogenesis. I will return to this relationship later. There was no relationship of genes in this category with *ANG* other than through shared functional annotation.

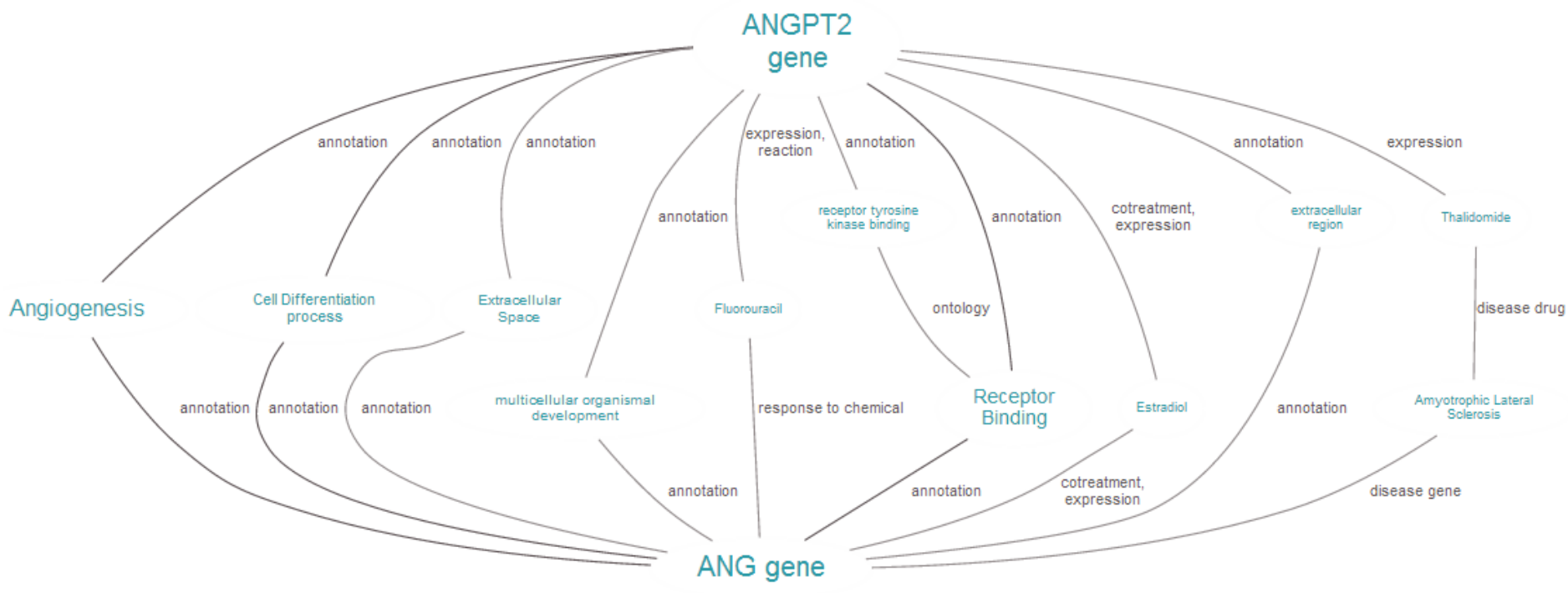


Figure 6-16. Interactions and pathways between genes *ANGPT2* and *ANG*

Angiogenesis and blood vessel development work towards the growth and maturation of healthy blood vessels. Angiogenesis itself implicates cell migration, proliferation and tubular structural formation as possible means of pathological dysfunction[306]. Furthermore the *ANG* gene contribution to angiogenesis can come through four (currently known) pathways; through ribonucleolytic activity, through membrane binding and basement degradation, through signalling transduction, and through enhancement of nucleic rRNA transcription.

ANG may be important to ALS, not only because several genomic findings and several genes found to significantly differentially express support angiogenesis as a pathogenic pathway, but the functions in which *ANG* is related to , have all been independently implicated in ALS. These include RNA regulation[57, 89, 307, 308], both membrane trafficking and binding[309], signal transduction[310, 311], and dysfunctional nucleic localisation[56]. Furthermore, angiogenin normally operates in endothelial which are neuroprotective for spinal motor neurons[306].

There are several significant functional annotation categories in the analysis of pathologically Mild-Moderate regions that lend support to the idea of angiogenesis and *ANG*-like dysfunction. Membrane activity, phosphatase activity (indicating problems with signal transduction) and EGF-like domains (opening up possible issues with cell growth and proliferation), were all highly enriched significant annotation clusters identified by DAVID. Examining the gene-expression cluster analysis for dose-dependent effects, we see genes *TMEM132E* (possible transmembrane protein) and HAMP (a signalling molecule for iron homeostasis) reduce in gene expression as the disease worsens, whereas gene expression for *ZKSAN1* (transcription regulation) and *REM1* (endothelial cell sprouting and cytoskeletal reorganization) gradually becomes increased when the disease is in its final stages, the last stage being death.

With four out of 14 functionally annotated clusters implicating gene expression abnormalities in angiogenin-like malfunction, with several functionally related genes showing graded change as the disease progresses, these results suggest the that angiogenin expression is pivotal in ALS pathogenesis.

The Mild-Moderate category was also extensively characterised by several functional annotated clusters involving glycoproteins, specifically membrane activity possibly linked to signalling and

Fibronectin. Although glycoproteins are ubiquitous and multifunctional, given the findings of the Mild-Moderate category, it is of interest that such proteins are most concentrated in blood. This may point to problem with membrane based signalling which leads to aberrant blood vessel development.

Glycoprotein functional annotated clusters in this analysis were the second and third most enriched. Fibronectin was the fourth most enriched, which is also a glycoprotein. Fibronectin is multifarious, involved in cell growth, adhesion, differentiation and migration. It was also first suspected to contribute to ALS pathology in 1980[312]. Increased levels of multi-domain glycoproteins in ALS tissue were found in 1992[313], and their presence have been replicated in various post-mortem material including cerebral spinal fluid[314] and plasma[315]. Glycoproteins have also showed differential expression in ALS mice[316], especially in motor neurons and astrocytes[264].

In this analysis and somewhat in ALS research there seems an unclear connection between angiogenin based activity and glycoprotein function. I found four genes (*SUSD1*, *CRIM1*, *KDR*, and *RAMP3*) previously implicated in ALS that showed a significant relationship with genes that differentially expressed in this analysis linking blood vessel development and glycoproteins through a wide variety of aberrant genetic functions.

So far I have covered eight of the 14 functional annotation clusters for the Mild-Moderate category. The remaining six annotated clusters point to abnormal expression concerning leukocyte migration in relation to cell motility, regulation of responses to an external stimulus, responses to a bacterial molecule, and chemotaxis. These alterations suggest a response to an environmental agent. The remaining two clusters relating to leukocyte activity were defence response, and surface antigen and binding, also suggesting an external agent.

Very few findings in ALS relate leukocyte activity as a whole to disease pathogenesis, however there is evidence of neuroinflammation and lymphocytes seem to play an important role. Before genetic advances in ALS began to accelerate around the mid 1990's, autoimmunity was a dominant hypothesis to why ALS occurred[317].

In patients there is a strong immunological response to the disease mostly in blood and the CNS[318]. IN ALS lymphocytes infiltrate perivascular cells (smooth muscle cells)[319] and invade the spinal cord in high concentrations. A study found 79% of ALS cases to have lymphocytic infiltration[320], in which the majority will likely be T-cell[321]. Lymphocytes show abnormal concentrations in patients early on in their disease, which suggests possible recruitment of T-cells to the rescue of dying motor neurons[322]. Neuroinflammation is mediated by T-cells as well as macrophages and mast cells[323], and this process has been shown to be neuroprotective towards motor neurons in mutation *SOD1* (*SOD1*^{MT}) mice[324] and humans through regulatory lymphocytes[325].

However, CNS T-cell responses can also be detrimental; for example this process slows Alzheimer's disease while it seems pathological in Parkinson's disease[326]. An ALS *SOD1*^{MT} mouse model manipulating the presence of lymphocytes showed them to be deleterious in the early stages of the disease, hastening onset through inhibition of M2 microglia[327]. The discrepancy between these consequences is likely to be due to the sub-classification of cell types and extent of ALS pathology. It is known that T lymphocytes can induce either cytotoxicity (M1 microglia) or neuroprotection (M2 microglia) in ALS, depending on disease stage[328].

What is also relevant here and important in this pathological pathway is that leukocytes interact with the Blood-Brain Barrier (BBB). In ALS, lymphocytes show small concentrations in the CNS until inflammation occurs which alter endothelium. This leads to expression of cellular adhesion molecules, which bind to leukocytes allowing leukocytes to cross the BBB into the CNS[329-332].

The evidence from the work presented in this chapter strongly points to leukocyte migration, cell motility, chemotaxis and surface antigen binding. All these functions are necessary for the leukocyte-BBB migration to occur. Furthermore blood vessel development and glycoprotein membrane signalling may also play a crucial role in allowing this to happen. For example, I found gene *IL7R* differentially expressing in both the second onset group in section 6.3.2.2 and the Severe pathology category in section 6.3.4.4. *IL7R* synthesises a glycoprotein that helps regulate lymphopoiesis. Taken together these findings closely link all three glycoprotein, angiogenin function, and leukocytes categories together.

In support of changes in immune-based gene expression, the cluster analysis and corresponding heat map of fold-change showed at least six immunogenes (*HLA-DRB1, IRF5, IL17RB, ZBTB7B, IL6, and CCL2*) that highly express in ALS cases to begin with and become unexpressed by the time the spinal region has reached Severe pathology. There are no immunogenes that increase expression over this period.

This analysis supports the involvement of immune function in ALS spinal cord pathology. The gene expression analysis of Mild-Moderate pathology can be divided into three major components, (a) angiogenesis and blood vessel development function, (b) glycoprotein-based membrane activity, and (c) leukocyte response. Membrane signalling and leukocyte responses are important for angiogenin efficacy, and all three are known ALS pathological pathways independently. Here I also suggest the idea that they are related, with leukocyte-led cytotoxicity occurring as the primary defect in one or all three of the above major components.

It is possible that a defect somewhere in the angiogenin or glycoprotein pathways cause the infiltrated CNS immune response which ultimately fails and becomes toxic. The fact that ALS research has found mutations in genes relevant to angiogenin and glycoproteins supports this hypothesis, whereas evidence for genetic involvement in immune based activity is non-existent. Furthermore it seems that the immunoresponse in the CNS, at least in terms of lymphocytes and neuroinflammation, can be beneficial if activated at the right stage of disease.

An alternative is that the pathology may arise in immunity itself which allows or leads to alterations in angiogenin and glycoprotein function. Under this reasoning we may conclude that genetic mutations found in genes like *VEGFA* and *ANG* are more like susceptibility factors rather than causers of ALS. In this scenario it is likely that immune responses are not caused by genetics as there is little evidence for this. The data here suggests that the immune response is towards an external stimulus or even pathogenic bacteria. But despite much research in ALS no convincing pathogen or environmental toxin has yet been identified.

A third hypothesis is that gene expression analyses here are all reacting to an unidentified pathology, of genetic or environmental origin. This is possible but unlikely as the findings replicate genetic and proteomic research that has previously shown direct causation between the functions described in this chapter and ALS. Furthermore, their implication in ALS pathology is supported by the bioinformatic post-hoc analyses showing interactions of differentially genes found here with known ALS mutated genes.

6.4.2.2 Moderate pathology

Moving on to the next category, Moderate pathology, we find the story changes little. No functional annotation cluster passed the 1.3 enrichment benchmark, but the most significant cluster was glycosylation and glycoproteins with an enrichment score of 1.21. What is hidden from this analysis is that several genes differentially expressing in this category are immunoresponse genes and show dose-dependent changes as the disease progresses; *HLA-C* expression increased initially with increasing severity and fell as the disease took hold, *BTN3A2* decreased with worsening pathology, whereas *PADI4* and *HLA-A29.1* increased expression.

The most significant differentially expressing gene, shown to greatly under-express and which has been implicated in every gene expression analysis of this thesis, was the glutamate transport gene *SLC1A7*. Please see section 6.3.1 for a detailed exploration of this gene in ALS.

The fourth most significant gene in this category, *GEMIN5*, physically interacts with *SMN1* to form macromolecular complexes involved in cytoplasmic assembly of snRNPs. It is also involved in nucleic pre-mRNA splicing. *SMN1* was first implicated in ALS through the examination of SMN2, in which abnormal copy numbers of *SMN1* was found to increase ALS susceptibility[101, 333, 334]. Now both SMN genes have been found to contribute to ALS pathology, which share a gene-gene relation where SMN2 counteracts the phenotypic expression of aberrant *SMN1*[335]. *SMN1* is thought to have a protective effect against *SOD1* toxicity[336]. Abnormal copy number of *SMN1* was initially found to be associated with ALS in a French ALS population[101] and interference with SMN

protein production shortens survival in mice[337]. It is also possible that *SMN1* shares a similar function to the ALS gene Fused in Sarcoma (FUS), with both genes perturbing cytoplasmic formation of snRNPs. How *GEMIN5* and *SMN1* relate to the current analysis is unclear other than possible link between transcription regulation and angiogenin function.

6.4.2.3 Moderate-Severe pathology

The most enriched and only significant functional annotation cluster for this category was metabolic and biosynthetic processes. This cluster contained ontological terms such as sulphur, coenzyme, and cofactor metabolic and biosynthetic processes. Increased concentrations of the sulphuric amino acid Taurine has been recorded in ALS[338, 339] and has some unclear relation to excitotoxicity[340]. Recently Taurine Transporter (TauT) has been shown to over express in *SOD1*^{MT} mice. It is likely that sulphur amino acids bind with glutamate receptors due to high-affinity[341], reducing receptor efficacy. Alternatively, the increase of sulphur amino acids in ALS may be a consequence of pathologically-related caspase enzyme activity.

The second most enriched functional annotation cluster was haemopoiesis and lymphoid organ development, and had an enrichment score of 1.18. These functions have not been directly linked to ALS but haemopoiesis is a synthesising process for components of blood cells and continues the theme found with early stages of the pathological severity.

6.4.2.4 Severe pathology

The final Severe pathology category somewhat describes a region of the spinal that has become damaged and fragmented through self-destruction. The most significant gene differentially expressing is *ANGPT2*, which has as one of its major functions endothelial cell apoptosis and vascular regression in the absence of *VEGFA*. Reduced *VEGFA* expression in *SOD1*^{MT} mice

leads to a more severe ALS phenotype[109] and reduces Age Of Onset (AOO) in human patients[342]. Increasing *VEGFA* expression led to greater protection of motor neurons from ischaemic death[109]. It has been proposed that *VEGFA* activates the PI3-K/Akt signal transduction pathway which reduces glutamate excitotoxicity[343]. I did not find any genes related to this pathway in this analysis. My findings suggest that the role of *VEGFA* in increasing motor degeneration severity is through *ANGPT2*, which will cause apoptotic death in endothelial cells when not modulated.

The Severe pathology category also found the gene *DNAH2* to be differentially expressed between cases and controls. *DNAH2* is a respiratory cilia protein associated with microtubules. It may be involved in ALS pathogenesis through Dynein and the *DCTN1* gene. *DCTN1* protein product Dynactin, when mutated, is believed to lead to problems of microtubule binding which impedes axonal transport. There is no immediate reason why this gene is present at the most Severe category of the disease where there should be no surviving motor neurons.

The top functional annotation clusters found for the Severe pathology category were striated muscle cell development, microsomes and vesicular fraction, and fatty acid metabolic processes (see Appendix Table A16). Paralysis of striated muscles is a consequence of denervation of motor neurons, which facilitates atrophy. It is possible that the expression activity found in the cluster represents an attempt to repopulate muscle cells in atrophied muscle. Fatty acid metabolites are too an indicator of cell death and neuroinflammation. Microglia synthesise fatty acid metabolites when producing inflammation and neurotoxic factors. It is most likely that the expression found in these two categories is from glia cells as they manage a post-apoptotic region.

Microsomes are usually cytoplasmic artefacts constituted from broken up endoplasmic reticulum (ER) organelle or plasma membrane and are vesicle-sized. Similarly, vesicular fraction refers to obsolete broken vesicular particles. These can occur via experimental procedures of RNA extraction and centrifugation, or they can occur naturally due to pathology. As (A) I do not see these functional annotation clusters anywhere else in the dataset, (B) the samples' RNA extraction was completely randomised and (C) extraction methods, laboratory, and scientist was the same, I will consider the microsomes to be naturally occurring. Furthermore it is what one

may expect to identify at the worse extremity of ALS pathology. Microsomes and vesicular fractions, in addition to apoptotic glia cell activity, are likely to be conclusion of the disease when it has depleted a spinal region of its motor neurons.

The gene-expression cluster-analysis and related heat-map concerning fold-change across each pathological category picked out several interesting changes in gene expression. Gene expression for *ACSF2* (catalyst of fatty acid metabolism) and *QDPR* (recycling enzyme) decreased as ALS worsened. *FBXL16* also decreased across categories. This is involved in the ubiquination of cell-cycle proteins and may also cause ubiquitin-like functions of protein degradation. Genes *STON2* (possible vesicle recycling), *CYP4F3* (contributes to a leukocyte chemoattractant), and *SFTPD* (an immune system collectin), revealed increased expression over the course of the disease.

These selected genes and the evidence presented above support a picture of a late, possibly toxic, immune response most likely from glia, as well as possible fragmented remnants of dead motor neurons.

6.4.3 Expression by severity category

Analyses of gene expression by different levels of tissue severity was performed to identify dose-dependent changes, as a proxy to changes in expression as the disease progresses through the spinal cord. Two methods were employed, the first being a genome-wide one-way ANOVA of all gene expression by severity (this was represented using a cluster analysis and heatmap), and the second method using the MFuzz soft-clustering algorithm which clusters genes based their on changes in expression as a consequence of severity.

The first method identified 63 genes with their expression significantly changing as a consequence of severity, with a non-adjusted p-value < 0.001. Two of these genes have previously been identified in ALS, *DIAPH3{Daoud, 2011 #1378}* and *VEGFA{Oosthuyse, 2001 #382}{Lambrechts, 2003 #590}*. Dauod et al. (2011) identified DIAPH3 through a candidate gene resequencing approach where they identified two missense mutations and one nonsense mutation. One of the missense mutation was predicted to affect protein function, with the nonsense mutation causing a premature stop codon.

VEGFA has previously been described in sections 6.4.2.1. Mild-moderate pathology and 6.4.2.4 Severe pathology. It was identified to interact with the highly significantly down-regulated gene ANGPT2. The one-way ANOVA directly implicates VEGFA as it clearly changed from up-regulation to down-regulation, when compared with its overall expression level, while moving from mild-moderately affected areas to severe areas. It seems that when ANGPT2 becomes down-regulated as does VEGFA at the same stage of the disease. A pathological mechanism may be that lack VEGFA excitatory activation of ANGPT2 leads to lack an endothelial cell proliferation. Or, looking at the relationship in converse, improper ANGPT2 activity leads to poor inhibition of VEGFA and therefore improper activation of apoptosis function.

The two clusters identified from the 63 genes identified by the ANOVA analysis were not significantly enriched. These were (1) intracellular protein transport and localisation, and (2) leukocyte response. The first cluster has relevancy in ALS pathology due to protein cytoplasmic mislocalisation of putative ALS-proteins as well as protein transport believed to be disrupted. Of interest in this analysis is how expression affected leukocyte response changes. It moves from down-regulation to up-regulation as the tissue is more severe. This concurs with the ALS

hypothesis regarding possible over-infiltration of C.N.S leukocyte that eventually become toxi to motor neurons.

The second method employed was to cluster genes based on their expression changes alongside varying tissue severity. The analysis identified five clusters, all enriched over 1.3 by DAVID's functional annotation analysis. Again we see an up-down change in leukocyte response but also coupled with regulatory transcription activity, suggesting that these two functions may be related. Interesting, but possibly ALS-relevant categories also included regulation of GTPase-activity, nucleic methylation, calcium-related cell adhesion, and nucleic lumen function.

6.4.4 Summary

This chapter approached the study of disease spread in ALS using gene expression analyses, histological analyses to estimate extent of pathology, and clinical history to establish where and when the disease began and spread, and how this changed in terms of gene expression.

My first approach was to use upper-limb cases only, where I used pathology and clinical data to establish where the disease began, showing that in these cases it began in the cervical region and moved rostro-caudally before finally involving the medulla. I then performed differential expression analysis between spinal segments of the same region, classifying each in sequence of occurrence in disease progression. Relatively few genes were found to differentially express and there were not enough genes to run gene-expression cluster analyses and functional annotation clustering. One gene *SLC1A7*, which synthesises a glutamate transporter called EAAT5, was found consistently to under-express throughout the spinal cord. This is a potential candidate for ALS pathogenesis as we know that *EAAT2* has been shown to be pathogenic in ALS.

My second approach of analysing spread was to divide pathological severity into four categories. I compared each category of ALS severity with controls. These analyses revealed a large number

of significantly differentially expressed genes which I was able to functionally cluster by annotation, and through gene-expression cluster-analysis identify genes that showed dose-dependent changes according to pathological severity. The overall picture revealed three main clusters of functions showing differential expression: (a) angiogenesis and blood vessel development function, (b) glycoprotein-based membrane activity, and (c) leukocyte response. As the analyses progressed through increasing levels of severity categories, clusters (a) and (b) became less prominent whereas immunity responses remained important. In light of other research there could be an inherent fault within any one of these systems, which eventually allows leukocytes to infiltrate the CNS possibly leading to toxicity or at least failure in protecting motor neurons.

My final approach was to identify which gene's expression changed in relation to the severity category of the tissue. My aim was to capture the change in gene function as ALS progresses. I identified two genes previously implicated in ALS, with one gene (VEGFA) showing a possible disrupted interaction with ANGPT2, which could be fatal to a cell. In addition, the cluster analysis of the ANOVA results statistically classified genes showing an up-down change in leukocyte response as the disease became more severe. This concurred with the results from my second approach where leukocyte response remained statistically present throughout each stage of the disease, even when the angiogenesis and glycoproteins functional categories started to waver as we move to more severely affected tissue. This third approach also identified clusters that are relevant to ALS, such as transcript variation and calcium-led cell adhesion.

Chapter 7 Temporal related disease phenotypes and gene discovery

7.1 Introduction

Disease genes in ALS have been categorised into *susceptibility* and *disease-modifying*. This chapter explores disease genes that modify ALS disease phenotypes. The major ALS phenotypes, which have been purported to have a genetic basis, some having heritability estimates to support this claim[132], are age of onset (AOO), duration from onset to death (survival), and disease onset site (bulbar, limb, or respiratory onset).

For an in-depth review of disease-modifying genes in ALS please see the literature review of this thesis. It is evident that disease-modifying genes are less frequently discovered than susceptibility genes, and have a poor history of replication. It may be that phenotypic data derived from multifarious clinics is not collected in a uniform manner, or that this area is under researched in ALS, or that there are simply fewer disease-modifying genes. A fourth possibility exists that the distinction between susceptibility and modification is fallacious and that all genes are both.

This chapter uses two analyses that aim to (a) calculate the variance explained by a UK genome-wide genotype dataset for AOO and survival, (b) identify genes using genome-wide linear regression that affect AOO and survival when modelled as quantitative traits. I also use logistic regression to examine onset site, dividing the population into three groups (bulbar, limb, and respiratory).

7.2 Methods

Sample collection, *sample preparation*, *genotyping*, *statistical quality control*, and *imputation* have been described previously (please see chapters 1 and 2). Only the UK samples used in Shatunov

et al. were used for this chapter's analyses because I did not have phenotype data for the other countries.

7.2.1 Advanced Complex Trait Analysis

Advanced Complex Trait Analysis (ACTA) is a modified version of the Genome-wide Complex Trait Analysis (GCTA) software[344, 345]. Both aim to calculate the amount phenotypic variance in a quantitative trait explained in a genome-wide SNP dataset, while also estimating "missing heritability" (variance in heritability of the trait not accounted for by the SNP dataset).

ACTA uses a mixed linear model (MLM); modelling the phenotypic variance as the dependent variable, fixed effects modelled with disease-modifying factors (in ALS, such as age, sex and population components), and SNPs modelled as random-effects. Modelling SNPs as random-effects allows their overall variance to be estimated and used to calculate how much the genetic variance accounts for the phenotypic variance. Modelling SNPs as random-effects is also useful because they are often not useful when estimating of the effect of an individual SNP on a phenotype, because genotypes vary from person to person. Instead, we can neatly calculate the non-specific effects of all SNPs on a phenotype.

By matching individuals, the estimation of genetic and phenotypic variance forms the basis of the genetic relationship matrices (GRM). ACTA then uses a restricted maximum likelihood (REML) analysis to calculate parameters for the GRM estimates, and test their likelihood given the dataset.

We used only autosomal SNPs in our dataset with a Minor Allele Frequency (MAF) of 0.01 to estimate the GRM. Our data consisted mainly of common genotype markers which is unlikely to tag rarer variants with a MAF of 0.01. Lower than this may result in an overestimation of the model.

We then pruned the matrices by (a) removing individuals with estimated relatedness greater than 0.025 (the recommended default value) and (b) assuming possible causal loci as having the same allele frequency distribution as genotyped SNPs. We then performed principal component analyses (PCA), estimating the first twenty eigenvalues. The eigenvalues represent latent groups (or variables) within the dataset, which we can use to account for extraneous variance that may interfere with estimations of genetic-phenotypic variance. These groups were included in the REML model.

The REML output consisted of variance estimates and standard errors for each model parameter; genetic variance, residual, and phenotype variance. It estimated genetic to phenotype variance ratio, and also performed a likelihood ratio test from the MLE in which the p-value is derived.

This was performed for both AOO and survival. Please note for both analyses all patients are deceased.

7.2.2 Genome-wide linear regression of quantitative traits

Genome-wide linear regression has been described in chapter 2. In brief, here we considered AOO and survival as quantitative traits for 578 UK cases. We quality filtered our dataset excluding SNPs with MAF <0.02, HWE <0.0001, and genotype genome-wide coverage of <95%. We also accounted for stratification effects caused by sex and adjusted for multiple testing.

830 SNPs were removed bringing the amount of SNPs analysed and corrected for 444,710, with a genomic inflation factor of 1.009. The inflation factor tests for unwanted population stratification by comparing the median observed distribution with the median expected distribution, using markers not relevant to the trait in question. The comparison is performed using chi-square.

I examined normality for both quantitative traits using Kolmogorov-Smirnoff in Stata 12.0. Both significantly deviated and AOO was transformed using the square of each data point, and for

survival the log. Re-testing normality showed the data distributions were no longer significantly different from normal.

I did not use genome-wide cox regression for survival data as patients were all deceased and so censor status is monomorphic, which is not supported by the R package GenABEL[346]. Cox regression survival analysis were run for individual SNPs highlighted as near significant in the survival genome-wide linear regression analysis. SNP alleles were used as covariates with survival in decimalised years and status, which all patients were deceased.

7.2.3 Genome-wide logistic regression of onset-site

Genome-wide logistic regression has been described in chapter 2. We quality filtered our dataset excluding SNPs with MAF <0.02, HWE <0.0001, and genotype genome-wide coverage of <95%. We also accounted for sex effects and adjusted for multiple testing for the 444,765 SNPs analysed.

There were 364 limb-onset, 163 bulbar-onset and 10 respiratory cases, and 4142 controls. All tests had a genomic inflation factor between 1.00 and 1.01.

7.3 Results

7.3.1 ACTA

We analysed AOO using ACTA (n = 572). The ratio of phenotype variance explained by our genotype data was 0.000001. A likelihood ratio test, examining whether this amount of variance explained was significant compared to a null model with no genetic variance component, was unsurprisingly not significant.

For AOO in the international dataset (n = 1366) this analysis was nearly significant with the genotype to phenotype variance ratio being 0.42 and a likelihood ratio test is 1.862 (p = 0.086). This analysis was conducting correcting for 20 eigenvectors.

Analysing our survival data (n = 572), the genotype to phenotype ratio variance was almost 1. The likelihood ratio test (2.96) was significant, with a p-value of 0.04. This suggests that a significant proportion of survival variance is explained by the genotype data.

7.3.2 Genome-wide linear regression

After quality control procedures there were 444,710 SNPs and the Bonferroni-corrected significance threshold was 1.12×10^{-7}.

We treated AOO as a quantitative trait and analysed this phenotype using a genome-wide linear regression (n = 599). There was a significant association at locus 18q22.3 in the UK set, involving 11 intergenic SNPs (see Figure 7-1 and Table 7-1).

We sought to replicate this finding in several international datasets, independently and combined. These SNPs did not reach significance in any dataset.

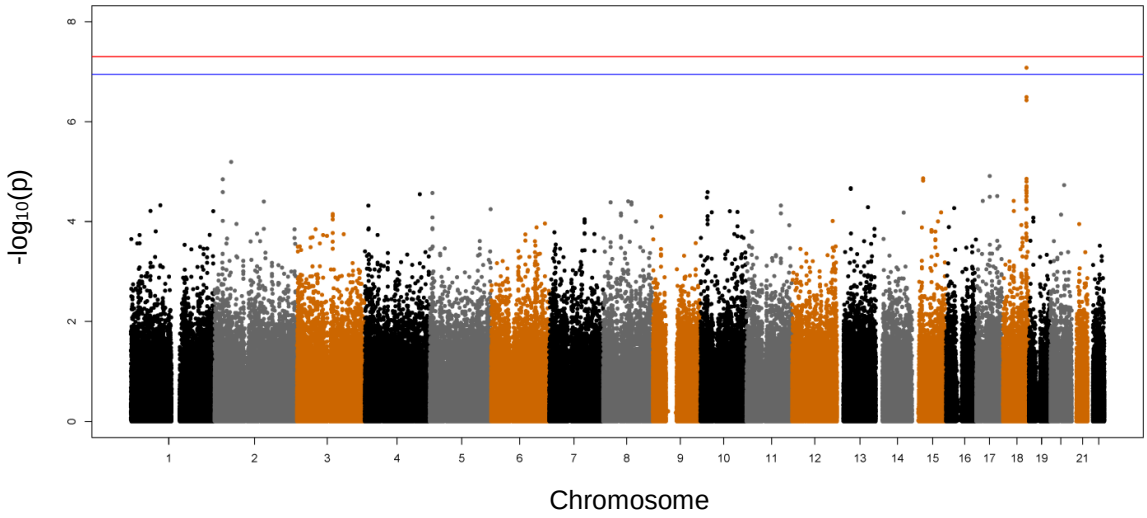


Figure 7-1. Manhattan plot of genome-wide linear regression of AOO

Because the 18q22.3 SNPs show a definitive association with ALS and because SNP rs10514080 is approximately 250kb from the nearest gene LOC100505817, I ran an epistasis analysis examining all significant chromosome 18 SNPs to try identify trans-acting interactions (Table 7-2).

Chr	SNP	Base Pos.	Allele	Beta	S.E.	P-value	Bonferroni Adjusted P
18	rs10514080	69426930	G	431	79.39	8.35E-08	0.03715
18	rs876964	69435617	A	396.3	76.66	3.23E-07	0.1435
18	rs1107827	69435868	A	394.8	76.79	3.74E-07	0.1662
2	rs4256004	48891658	G	-506.6	111.3	6.42E-06	1
17	rs4792941	39848546	A	-340.7	77.24	1.23E-05	1
15	rs7164838	32754866	A	359.5	81.96	1.37E-05	1
18	rs2850481	69445149	G	356	81.26	1.40E-05	1
2	rs2080727	24204411	G	337	77	1.43E-05	1
15	rs1978809	32745481	A	428.1	98.14	1.52E-05	1
18	rs1943902	69455480	A	353.1	81.11	1.59E-05	1
20	rs7272567	40219855	A	-647.6	150.1	1.87E-05	1
18	rs6566754	69479113	A	343.6	79.81	1.96E-05	1
18	rs11151895	69512617	A	338.1	78.79	2.08E-05	1
18	rs9950388	69415070	G	319.1	74.39	2.10E-05	1
13	rs2282022	40272186	G	330.8	77.24	2.16E-05	1
13	rs1536805	40267658	A	338.1	79.08	2.23E-05	1
13	rs3783163	40278453	G	338.1	79.08	2.23E-05	1
18	rs1484217	68610550	A	332.6	78.05	2.37E-05	1
18	rs8099013	69412772	A	317.5	74.63	2.45E-05	1
2	rs4149372	24159016	C	332.7	78.43	2.58E-05	1

Table 7-1. Genome wide linear regression of AOO showing top 20 most significant SNPs

Chr	SNP	No. of Tests	Proportion significant	Best χ^2	Best Chr	Best SNP
18	rs6566754	444785	0.009688	31.01	7	rs6947780
18	rs1943902	444785	0.009717	27.39	7	rs6947780
18	rs2850481	444785	0.00971	26.52	7	rs6947780
18	rs10514080	444785	0.009139	24.56	7	rs6947780
18	rs1484217	444785	0.01025	22.29	10	rs10761563
18	rs876964	444783	0.009234	21.66	10	rs10999801
18	rs11151895	444785	0.009382	21.42	21	rs2827206
18	rs1107827	444783	0.009097	21.26	10	rs10999801
18	rs9950388	444785	0.01028	20.31	3	rs11720065
18	rs8099013	444785	0.01048	19.86	19	rs9304656

Table 7-2. Genome wide epistasis analysis using significant chromosome 18 SNP

Four out of 10 SNPs interact with rs6947780, which is an intergenic SNP approximately equidistant from RNA gene *LINC00265* and *CDK13* by 100kb. Neither gene has been previously implicated in ALS.

rs10761563 on chromosome 10q21.2 is in intron 3 of the *RHOBTB1* gene, and its function is GTPase signal transduction and actin filaments regulation. rs10999801 is an intergenic SNP 10q22.1 and does not share linkage disequilibrium (LD) with rs10761563. rs2827206, rs11720065 and rs9304656 are all also intergenic.

We also treated survival as a quantitative trait in a genome-wide linear regression (n = 599), which identified no significantly associated SNPs after Bonferroni correction. There were two SNPs showing suggestive association that were in high LD with each at $r^2 = 0.98$ and D' = 0.99. These

were rs3781399 (unadjusted p = $9.62x10^{-06}$) and rs11818446 (unadjusted p = 1.32×10^{-5}), in gene CTBP2 on chromosome 10. CTBP2 has not previously been implicated in ALS.

Cox regression of SNP rs3781399 with alleles as covariates confirmed that it had a significant impact on ALS survival (p = 0.001, Exp (B) = 0.601). 48 patients had alleles AG and 529 had alleles GG (21 cases had no survival data and 1 case was exclude for being under-represented with alleles AA). The G allele is risk and significantly decreased survival in ALS (see Figure 7-2).

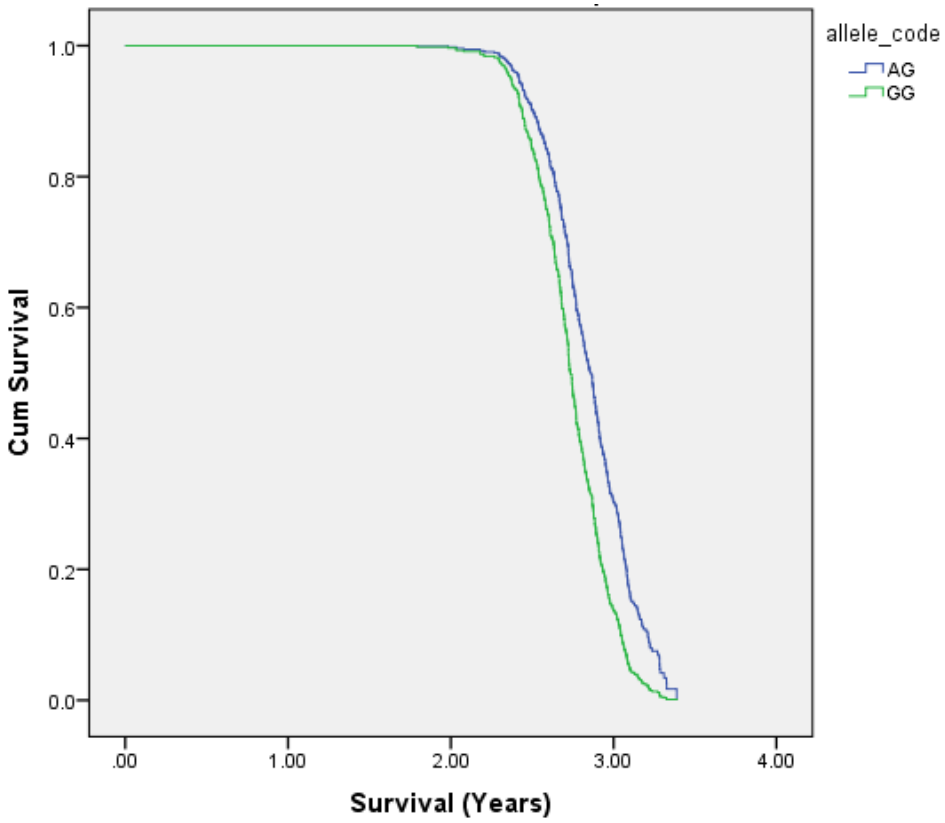


Figure 7-2. Survival plot of cox regression of rs3781399 alleles effect on ALS survival

Cox regression of SNP rs11818446 with alleles as covariates confirmed that it had a significant impact on ALS survival (p = 0.001, Exp (B) = 0.601). 49 patients had alleles GA and 528 had alleles AA (21 cases had no survival data and 1 case was exclude for being under-represented

with alleles GG). The A allele is risk and significantly decreased survival in ALS (see Figure 7-3).

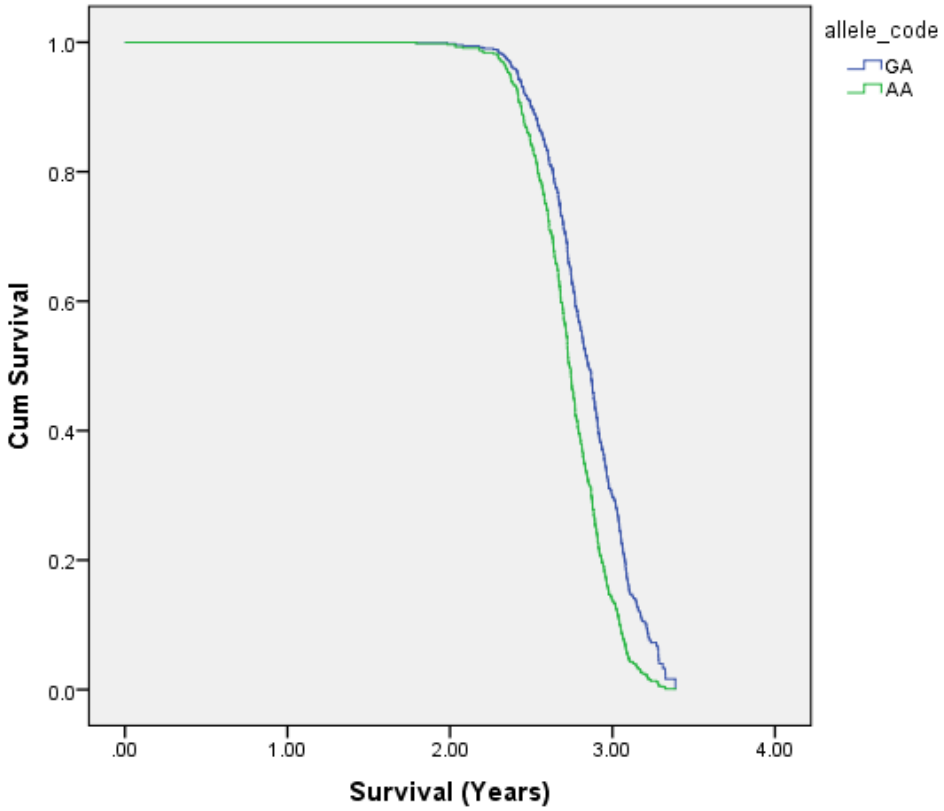


Figure 7-3. Survival plot of cox regression of rs11818446 alleles effect on ALS survival

7.3.3 Genome-wide logistic regression

7.3.3.1 Limb-onset

After quality control procedures there were 444,765 SNPs, and the Bonferroni-corrected significance threshold was 1.12×10^{-7}.

A genome-wide logistic regression association analysis was performed comparing limb-onset patients (n = 364) with controls (n = 4142). No SNPs were significantly associated after Bonferroni

correction. However four pairs of SNPs, sharing the same genomic loci, had p-values that were suggestive of a disease association. To elucidate possible disease relevant haplotypes for each SNP-pair, linkage disequilibrium was calculated across all cases and controls (Table 7-3).

Chr	SNP1	SNP2	Linkage Disequilibrium r^2	Linkage Disequilibrium D'
10	rs11192617	rs951030	1	1
10	rs8624	rs7904918	0.87	0.96
22	rs9615919	rs9615341	0.91	0.96
8	rs1494913	rs1389976	0.91	0.96

Table 7-3. Linkage disequilibrium for SNP-pairs showing association with limb-onset ALS

Chromosome 10 SNP rs8624 (unadjusted p = 1.09×10^{-5}) is exonic and rs7904918 (unadjusted p = 1.46×10^{-5}) is intronic in the gene *ACSL5*. Furthermore chromosome 22 SNP rs9615919 (unadjusted p = 2.16×10^{-5}) is exonic in *LOC284933* and intronic in *FAM19A5*, and rs9615341 (unadjusted p = 3.22×10^{-5}) is intronic in both *LOC284933* and *FAM19A5*. The remaining SNPs are intergenic and are not placed near a gene. FAM195A has been previously identified in ALS[128, 245].

When examining cases and controls under a logistic regression not stratified by onset-region, rs8624 decreases in association by an order of magnitude (to p = 3.99×10^{-4}) as does rs7904918 (to p= 5.67×10^{-4}), rs9615919 (to p= 7.72×10^{-4}) and rs9615341 (to p= 3.85×10^{-4}). It is therefore

possible that these SNPs associate with limb-onset ALS more frequently than ALS absent of onset location specification.

7.3.4 Bulbar-onset

After quality control procedures there were 444,765 SNPs and Bonferroni-correction significance threshold was 1.12×10^{-7}.

A genome-wide logistic regression association analysis was performed comparing bulbar-onset patients (n = 163) with controls (n = 4142). No SNPs were significantly associated after Bonferroni correction. However four pairs of SNPs, sharing the same genomic loci, showed a trend to association. To elucidate possible disease relevant haplotypes for each SNP-pair, linkage disequilibrium was calculated across all cases and controls as (Table 7-4).

Chr	SNP1	SNP2	Linkage Disequilibrium r^2	Linkage Disequilibrium D’
7	rs2049819	rs7805607	0.88	0.99
1	rs11162065	rs12081931	1	1
1	rs7411511	rs1441831	1	1
7	rs10259252	rs17160623	0.76	0.99

Table 7-4. Linkage disequilibrium for SNP-pairs showing association with bulbar-onset ALS

Chromosome 7 SNPs rs10259252 (unadjusted p = $1.09x10^{-5}$) and rs17160623 (unadjusted p = $1.09x10^{-5}$) are both intronic of the gene *KIAA1549*, which has not been implicated in ALS.

When examining cases and controls under a logistic regression not stratifying by onset-region, rs10259252 decreased in its association with ALS by three orders of magnitude (from p = $1.17x10^{-2}$) and rs17160623 by two orders of magnitude (from p = $7.36x10^{-3}$). This presents the possibility that for this bulbar-onset cohort, these SNPs are important.

7.4 Discussion

Calculating the proportion of UK AOO phenotype variance explained by our genotype data, using ACTA, estimated that there is little genetic contribution to AOO in ALS. Somewhat contradictory is that AOO variance in the international dataset showed a near-significant proportion explained by its genotype data. The proportion of UK survival phenotype variance was significantly explained by our genotype data. We did not have international phenotype data.

There was an intergenic SNP on chromosome 18q22.3, rs10514080, which significantly predicted AOO when modelled as a quantitative trait linear regression. Furthermore, 9 of the top 20 SNPs significantly predicting ALS AOO in this region showed LD with one another suggesting several disease related haplotypes. A genome-wide epistasis analysis of the 18q22.3 SNPS shows in four instances a significant interaction coefficient with rs6947780.

The epistasis analysis did not elucidate any non-intergenic SNPs and it is not immediately clear what the biological basis may be for the 18q22.3 locus, and rs6947780, influencing AOO in ALS. One SNP was found to be intronic in the gene *RHOBTB1*, a gene involved in GTPase signal transduction and actin filaments regulation. The gene has not been previously implicated in ALS, however signal transduction[347, 348] and actin[349] have had been implicated in the ALS proteomic or

pathology research previously. Actin is central to *PFN1* functionality, a gene recently implicated in ALS[235], as is *KIFAP3* implicated in survival, and *NEFH* showing association with ALS and coding for an essential component of the neuronal cytoskeleton.

Linear regression of the survival phenotype revealed two SNPs in LD, with suggestive p-values but not beyond the Bonferroni adjusted correction, in gene *CTBP2* on chromosome 10q26.2. A gene not previously implicated in ALS before, it codes for a protein involved in synapse specialisation and as a transcription repressor. We did not have an international dataset containing survival to replicate this finding.

I also performed logistic regression, dividing the UK case population into limb and bulbar-onset. No SNP was significant beyond Bonferroni correction, but pairs of top-SNPs within the top 10 most significant associations were highlighted as interesting. All pairs had LD greater than r^2 = 0.8. The genes *ACSL5*, *FAM19A5* with sub-gene *LOC284933*, were implicated as associated with limb-inset using this method. *ACSL5* is involved cell growth in fatty acids, a function that I was not able to identify in previous ALS research. Interestingly *FAM19A5* and *LOC284933* were identified in Chapter 2 as having an epistasis interaction with the hexanucleotide repeat mutation. It is unknown why these SNPs increase in significance for limb-onset and have an epistatic interaction with the hexanucleotide repeat mutation, as *C9ORF72* carriers are known to have both bulbar and limb onset[350].

For bulbar-onset *KIAA1549* was a candidate. Unfortunately not much is known about the function of this gene. It has been implicated in gliomas and known to transcriptionally fuse with BRAF, an oncogene[351].

Weaknesses of the above analyses are that ACTA most likely required a larger dataset than 572 cases. It is probable that is why AOO phenotype variance explained genotypic variance was near significant in the international dataset. It would be contradictory to existing research which has identified genes that directly and indirectly influence AOO in ALS[137, 352]. Furthermore we were able to identify a significant SNP predictor of AOO in ALS; although the two methods are fundamentally different and ACTA does not identify specific SNPs.

Power to detect SNPs that significantly predict phenotype variance in the UK dataset was an issue, but there many SNPs in the top 10 shared LD with another top 10 SNP. Although these LD relationships do not discount the possibility of false positives they do make them less likely.

In relation to previous findings, the linear regression analysis did not replicate SNPs within genes known to affect AOO in ALS, including SNPs from the recently associated 1p34.1 loci {Consortium, 2013 #1116}, beyond bonferroni correction. One SNP rs17202060 ($p = 0.009$) showed prediction without multiple-testing corrections. This SNP lies in intron 13 of the previously implicated gene TXNRD1. Because of the large variance of AOO between patients, likely to multiple genetic and non-genetic factors, it is likely their identification requires much statistical power. AOO variance caused by genotype variance became only nearly significant when including the four countries used in this analysis. These results indicate that the genetic influence on AOO is far less than the genetic influence of survival.

Similarly, I did not replicate genes found to modulate survival in ALS after multiple testing adjustment. A SNP lying in intron 8 of KIFAP3 showed prediction of ALS survival with $p = 0.038$.KIFAP3 has been previously implicated in predicting ALS survival time. In contrast to AOO, the genotype:survival variance ratio was very significant, even within the UK only. Survival seems to be influenced by multiple small effect-size mutations, unidentifiable using genotype association methods.

In summary, I found several genes that may be interesting to ALS pathology, when stratifying GWAS analyses by phenotype.

Chapter 8 Health utility and ALS clinical stage

8.1 Introduction

ALS is relentlessly progressive with over 50% of patients dying in the first 30 months, and only 20% surviving more than 5 years after symptom onset[353]. As a result, it is a greatly feared diagnosis and the commonest reason to seek assisted suicide[354]. Although the poor prognosis means that ALS appears rare, the lifetime risk is one in 300 by age 85 years[355], which represents a significant economic burden. A study of the cost of ALS care in the Netherlands estimated the average cost per patient at €1336 per month, with cost increasing significantly towards the end of life[356]. A US study found that inpatient mechanical ventilation and nursing contributed considerably to the $300,000/annum cost of patients in the terminal phase of their disease[357].

Advances in the understanding of ALS, and its pathogenesis, management and treatment have largely occurred in the past decade[358]. Clinical trials have increased in frequency as national and international ALS consortia have formed, and novel therapeutic targets discovered. Many clinical trials use functional rating scales, and generic or disease-specific health-related quality of life (HRQoL) questionnaires to measure primary and secondary outcomes. A commonly used HRQoL measurement, recommended by the National Institute for Health and Care Excellence (NICE) for cost-effectiveness analysis[359] is the EQ-5D, developed and validated for the UK population by the EuroQol Group[360].

We recently proposed a clinical staging system for ALS, defined by milestones in disease progression occurring at predictable proportions of time elapsed between onset and death[163].

A major use of clinical staging systems is in clinical trials, either as a trial endpoint, or to allow an economic analysis by assessing time spent in each health state, with health states mapped to utility. Few disease-specific quality-of-life instruments have been used consistently to quantify health utility in ALS[361], and there is little knowledge of whether any ALS clinical staging system can also predict patient quality of life or other secondary outcomes, such as scores on the Hospital Anxiety and Depression Scale, HADS. We therefore explored whether the King's ALS staging system could predict EuroQol EQ-5D health utility and HADS scores using data from a large multicentre clinical trial of lithium carbonate in ALS (UKMND-LiCALS Study Group[362,363]) in which patients completed the EQ-5D, HADS and ALS Functional Rating Scale questionnaires at each visit.

8.2 Methods

8.2.1 Setting and patients

EQ-5D and HADS questionnaires were collected from participants in the Lithium Carbonate in ALS (LiCALS) 18 month, multi-centre, double blind, randomized parallel group trial[362,363].

Questionnaires were completed at the pre-randomization visit and prior to the start of study medication, and subsequently measured at: Months 3, 6, 9, 12, 15, and 18.

We also collected basic demographic and clinical information such as age, sex, site of disease onset, age at onset, and ALS phenotype.

8.2.2 Estimation of clinical stage

Clinical stage was not collected prospectively as part of this trial. We therefore used an algorithm based on responses to the ALS Functional Rating Scale-Revised (ALSFRS-R) to estimate clinical stage. This is possible because the domains in this system partly overlap the domains used to calculate stage.

8.2.3 Modifications to King's ALS Staging System

Stage 1 patients are pre-diagnosis, and so were not eligible to enter the LiCALS trial. As a result there are no EQ-5D or HADS questionnaires labelled Stage 1 in the analysis, but those at Stage 2A are the same as Stage 1 clinically, only differing in diagnostic status (Figure 8-1). Stages 4a and 4b can present in a varying order partly contingent on whether the patient had bulbar or limb onset of symptoms. While those with bulbar-onset usually require gastrostomy before non-invasive ventilation, those with limb-onset usually require non-invasive ventilation before gastrostomy. As they occur at the same proportion of time through the disease course, these two stages were combined and labelled as Stage 4 for this analysis (Figure 8-1).

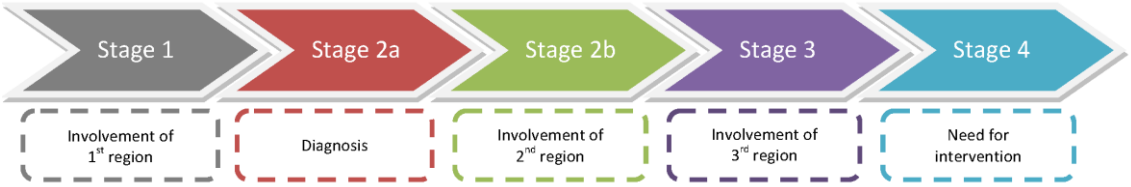


Figure 8-1. King's ALS Clinical Stages. Those at Stage 2a are the same as Stage 1 clinically, only differing in diagnostic status

8.2.4 Instruments

The EQ-5D is a generic health-related quality-of-life questionnaire that is primarily designed for self-completion by respondents. The questionnaire is split into two parts: the descriptive section and the visual analogue scale (VAS). The descriptive section classifies health status by measuring perception of functioning in five dimensions: Mobility (Mob), Self-care (SC), Usual activities (UA), Pain/discomfort (PD), and Anxiety/depression (AD). Each dimension is given a score out of three: no problems (1), some problems (2), and extreme problems (3). This results in 243 possible combinations. Through the use of a 'value set' (a set of weights), the five descriptive scores are then converted into a single aggregate 'health utility' score anchored at 1 (perfect health) and 0 (death). The EQ-5D value set used in LiCALS is the most widely used in the United Kingdom, and is known as the "Measurement and Valuation of Health A1 Value set". It is derived using time-trade off methods, and has been recommended by the Washington Panel on Cost Effectiveness in Health & Medicine and validated by EuroQol for use in cost-effectiveness analysis[364, 365].

The visual analogue scale component of the EQ-5D is a rating scale from 1 (worst imaginable health) to 100 (best imaginable health)[366]. The EQ-5D dimensions were also analysed individually, and were treated as continuous variables for better clinical interpretability. Each dimension resembled a normal distribution and statistical analysis corresponded to equivalent non-parametric multinomial regression analyses (not reported).

For the remainder of this article 'EQ-5D *health utility'* will refer to the aggregate EQ-5D utility score, and 'EQ-5D *vas*' will refer to the score given by a patient on the VAS.

The HADS self- assessment form screens for anxiety and depression via an aggregated sum from seven questions. It is designed specifically for medical practice and physical health settings and can be used to assess how symptom severity and prognosis influence anxiety and depression in patients. We used an ALS-specific revised HADS scale, in which a question is omitted from both depression and anxiety sub-scales, and the scoring structure of two items for depression is altered[367].

8.2.5 Statistical analysis

Statistical analysis packages Stata v12.0 and IBM SPSS v21.0 were used. To assess distribution normality we used a Shapiro-Wilks test and data were transformed to normality where necessary. To assess heteroscedasticity, the Cameron's and Trivedi's decomposition of IM-test was used.

We assessed four outcome measures individually (Health Utility, EQ-5D VAS, EQ-5D Dimensions, and HADS), each taken at seven different observations throughout the trial. We tested which ALS disease predictors significantly affected these outcome measures. Predictors examined were sex, Age of Onset (AOO), disease onset site, and treatment group (Placebo/Active), and the main predictor of interest: ALS clinical stage.

To test the statistical efficacy of the ALS clinical stages we created four models:

Model 1. *Null Model*: $y_{ij} = \mu + f_i + e_{ij}$

Model 2. *Stage-only Model*: $y_{ij} = \mu + \beta_1 x_{ij} + f_i + e_{ij}$

Model 3. *Non-stage covariates Model*: $y_{ij} = \mu + \beta_{2...5} x_{ij} + f_i + e_{ij}$

Model 4. With-stage covariates Model: $y_{ij} = \mu + \beta_1 x_{ij} + \beta_{2...5} x_{ij} + f_i + e_{ij}$

Where y = Outcome measure, μ = Grand Mean, f = between subject variation, e = Random error, $\beta_1 x$ = Clinical Stage predictor, $\beta_{2...5} x$ = Other ALS predictors, $_i$ = observation, and $_j$ = case

We tested the above models using random-effects regression with maximum likelihood estimation (MLE). This estimates the parameters of these models and informs us (except for Model 1) how well the predictors (for example AOO) predict the outcome measures (for example health utility). Random error caused by repeated measures across time for each case was accounted for by applying Person ID as the grouping variable, so we can test across case.

To compare models we used a log likelihood ratio test. We compared the *Stage-only Model* (2) versus the *Null Model* (1) to assess clinical stage as an individual predictor, and compared the *With-stage Covariates Model* (4) versus the *Without-stage Covariates Model (3)* to assess clinical stage as a predictor alongside typical ALS predictors. We calculated how much additional variance was accounted for by inclusion of clinical stage into a model.

To assess efficacy in identifying statistical differences in outcome measures between ALS clinical stages we coded each stage versus stage comparison using dummy variables within the MLE regression. We report coefficient and mean differences, z-test and related probability statistics.

To obtain *Cohen's D* we collapsed the eight observations, corresponding to trial visits, into the four clinical stages, and used the mean differences, standard deviations, and Pearson's correlation between stages. Where there were repeated measures at a particular stage we simply used the first observation. As a general rule, *Cohen's D* effect sizes are regarded as small at 0.2, medium at 0.5, and large at 0.8.

8.3 Results

8.3.1 Patient demographics

Table 8-1 shows the characteristics of the 214 individuals included in 10 centres participating in the LiCALS trial (for missing values see Appendix Table A20). The majority of patients were Caucasian (98%) and male (68%). The mean age of onset was 58 (s.d. ± 11 years) with 22% having bulbar-onset, and 78% of all cases were apparently sporadic. Ninety-seven patients had died by the end of the trial at month 18.

Characteristic	Summary statistics		Total n=214
Gender	**Females: n (%)**		66 (31%)
	Males: n (%)		148 (69%)
Ethnicity	**White: n (%)**		210 (98%)
Age at Onset	**mean (sd)**		58·1 (10·8)
Site of onset	**Bulbar: n (%)**		47 (22%)
Type of onset	**Sporadic: n (%)**		167 (78%)
El Escorial diagnostic category	**Clinically definite ALS: n (%)**		82 (38%)
	Clinically probable ALS: n (%)		80 (37%)
	Clinically probable - laboratory supported ALS: n (%)		38 (17%)
	Clinically possible ALS: n (%)		14 (5%)
Number of Individuals progressing through each King's ALS Stage	**Stage 2**	**2a**	75
		2b	125
	Stage 3		143
	Stage 4	**4a**	24
		4b	48

Table 8-1. Patient Characteristics

8.3.2 Health utility

The distribution of health utility scores was close to normality and so was not transformed. We assessed Model 1, comprising just the grand mean and random error, which gave a log likelihood score of -126.02. When we included clinical stage (Model 2) the log likelihood was -53.48. Comparing the two models using a likelihood ratio chi-square test we found that the with-stage model was significantly a better predictor of health utility scores ($\chi^2 = 145.08$, $p = 3.14 \times 10^{-32}$).

Using likelihood-ratio analyses, we found Model 3, which included sex, AOO, disease onset site, and treatment Group, was no better at predicting health utility than Model 1 (null model) ($\chi^2 = 2.24$, $p = 0.32$). Model 4 (which included clinical stages and all predictor covariates from Model 3) compared with Model 3 was a significantly better predictor of health utility ($\chi^2 = 146.45$, $p = 1.57 \times 10^{-32}$), confirming that clinical stage predicts health utility.

Clinical stages were coded as dummy variables to compare differences in health utility between stages using an MLE regression model. Differences in health utility were significant for comparisons of all clinical stages (Table 8-2), on average reducing by -0.12 points linearly through each consecutive stage, with significant differences between stage 2a and 2b ($Z(1) = -4.30$, $p = 1.7 \times 10^{-5}$), 2b and 3 ($Z(1) = -0.56$, $p = 6.8 \times 10^{-8}$), and stage 3 and 4 ($Z(1) = -5.56$, $p = 2.8 \times 10^{-8}$) (Figure 8-2).

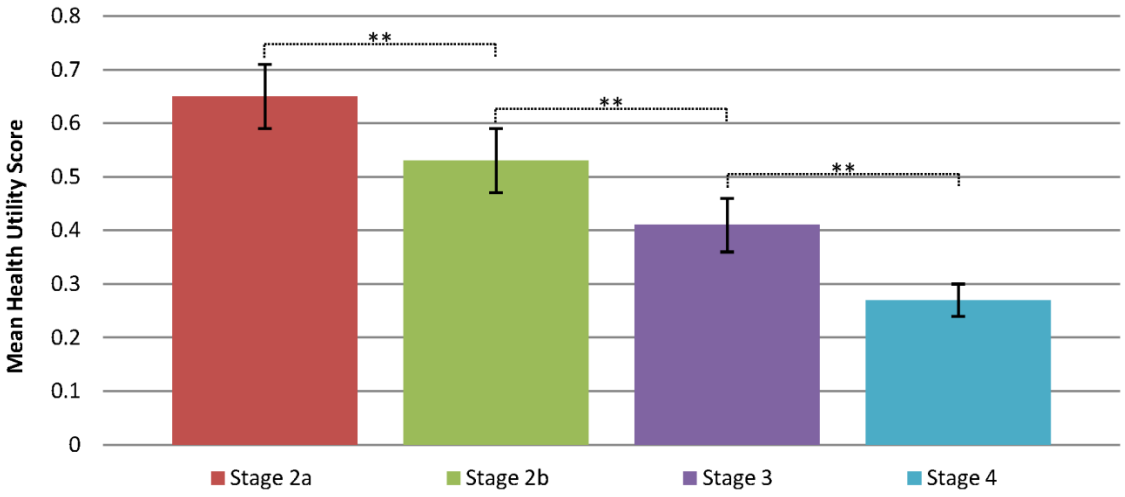


Figure 8-2. Mean health utility scores across each clinical stage, showing p-values as asterisks.

** $p < 0.01$

* $p < 0.05$

ALS stage	Utility			*Change* in utility when moving to ALS stage								
				Stage 2b			Stage 3			Stage 4		
	Mean	Mean		95% CI	Mean		Mean	95% CI		Mean	95% CI	
2a	0·65	0·59	0·71	-0·12**	-0·17	-0·06	-0·24**	-0·18	-0·30	-0·38**	-0·44	-0·32
2b	0·53	0·49	0·58				-0·12**	-0·17	-0·08	-0·26**	-0·32	-0·21
3	0·41	0·36	0·46							-0·14**	-0·19	-0·09
4	0·27	0·24	0·30									

Table 8-2. MLE regression comparing mean health utility scores between clinical stages.

** $p < 0.01$

* $p < 0.05$

8.3.3 Visual Analogue Scores

The distribution of VAS scores was close to normality and we did not transform the data. The *Null* Model 1 gave a log likelihood score of -3951.09, and Model 2 gave a log likelihood of -3920.30. Using a likelihood-ratio test comparing the two models we found Model 2 (which includes clinical stage) to be a statistically better predictor of VAS scores ($\chi^2 = 61.575$, $p = 4.21 \times 10^{-14}$).

We compared Model 3 with the *null* Model 1 using a likelihood-ratio test, and found no difference ($\chi^2 = 0.87$, $p = 0.65$). However, Model 4 was a significantly better predictor of VAS than Model 3, due to the inclusion of clinical stage ($\chi^2 = 62.37$, $p = 2.86 \times 10^{-14}$).

Coding clinical stages as dummy variables, we found differences between all clinical stages for mean VAS scores, except between stages 3 and 4 ($Z(1) = -1.75$, $p = 0.081$, *Cohen's D* = 0.50) (Table 8-3). The fact that the difference between stages 3 and 4 are near significant and has a medium effect size when stages are collapsed (n = 40), suggests that the non-significant p-value may be due to power issues and missing data points. Mean VAS scores fell -14.90 points between stage 2a and stage 4. Closer inspection of the VAS analysis between stage 2a and 2b showed that $Z(1) = -2.5$, $p = 0.013$ and between stage 2b and 3 that $Z(1) = -4.74$, $p = p = 2.2 \times 10^{-6}$ (Figure 8-3).

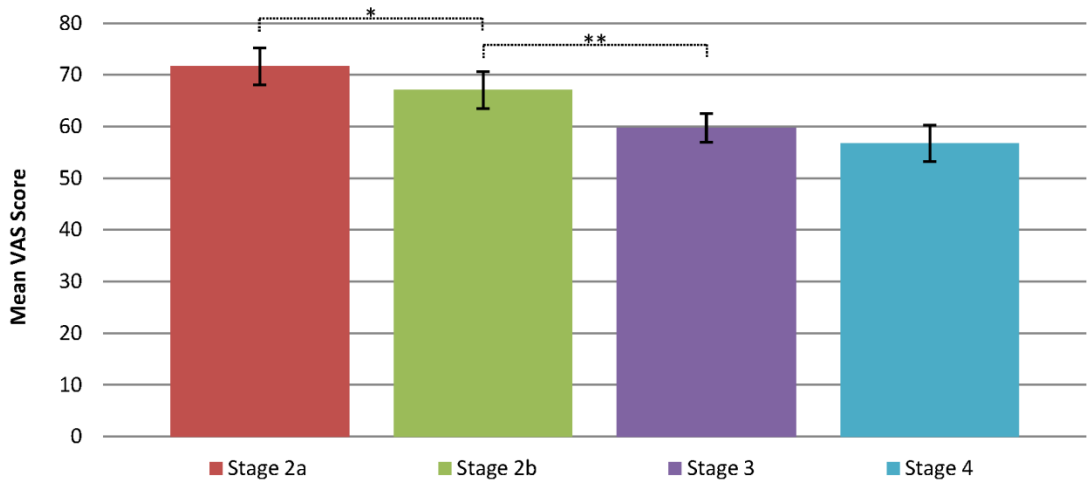


Figure 8-3. Mean VAS scores across each clinical stage, showing p-values as asterisks.

** $p < 0.01$

* $p < 0.05$

ALS stage	VAS			*Change* in VAS score when moving to ALS stage								
				Stage 2b			Stage 3			Stage 4		
	Mean	95% CI		Mean	95% CI		Mean	95% CI		Mean	95% CI	
2a	71·65	68·05	75·24	-4·59**	-8·18	-0·99	-11·89**	-15·64	-8·14	-14·90**	-19·11	-10·69
2b	67·07	64·09	70·04				-7·31**	-10·33	-4·28	-10·32**	-13·95	-6·68
3	59·76	56·97	62·55							-3·01**	-6·39	-0·37
4	56·75	53·23	60·27									

Table 8-3. MLE regression comparing Mean VAS scores between clinical stages.

** $p < 0.01$

* $p < 0.05$

8.3.4 EQ-5D dimensions

Clinical stage was a significant predictor for each EQ-5D dimension comparing Model 2 against Model 1: Mob (χ^2 = 105.02, p = 1.66×10^{-23}), SC (χ^2 = 123.69, p = 1.39×10^{-27}), UA (χ^2 = 85.77, p = 2.39×10^{-19}), PD (χ^2 = 18.08, p = 1.19×10^{-4}), and AD (χ^2 = 15.33, p = 4.70×10^{-4}). This was even more so comparing Model 4 with Model 3: Mob (χ^2 = 105.35., p = 1.33×10^{-23}), SC (χ^2 = 124.38, p = 9.78×10^{-28}), UA (χ^2 =87.21, p = 1.16×10^{-19}), PD (χ^2 = 17.84, p = 1.34×10^{-4}), and AD (χ^2 = 14.98, p = 5.58×10^{-4}).

Treating EQ-5D dimension scores as continuous variables we tested differences between all stages and the subsequent stage (see Appendix Table A21; Figure 8-4). We identified significant differences between all clinical stages for Mobility, Self-Care, and Usual Activity. For Pain/Discomfort significant differences were not found between stages 2b and 3 (Z(1) = 1.26, p = 0.21), 2b and 4 (Z(1) = 0.39, p = 0.70), and 3 and 4 (Z(1) = -0.71, p = 0.48). For Anxiety/Depression a significant difference was not found between stages 2b and 3 (Z(1) = -0.39, p = 0.69). These results are consistent with what is known about ALS progression.

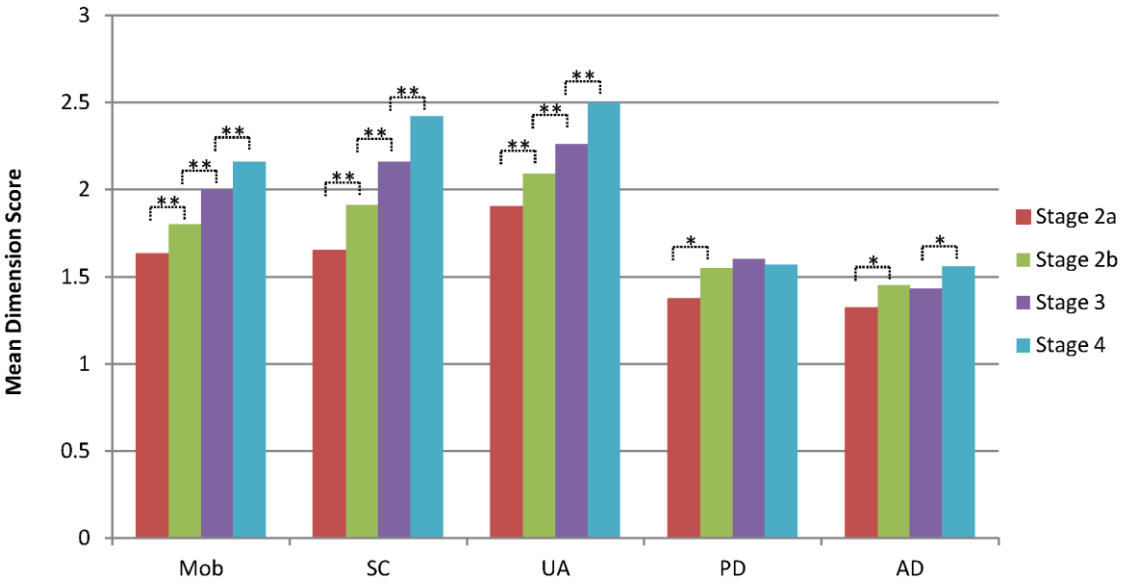


Figure 8-4. Mean EQ-5D dimension scores by dimension stratified by clinical stage, showing p-values as asterisks.

** p < 0.01

* p < 0.05

8.3.5 Hospital Anxiety and Depression Scale – Depression

The distribution of both HADS depression and anxiety scores were slightly skewed to lower scores but no transformation significantly improved the distribution. The log likelihood score for *Null* Model 1 was -1879.92 and for Model 2 was -1858.11. Model 2 was a significantly better predictor of HADS depression scores with clinical stage ($\chi^2 = 37.62$, $p = 6.78 \times 10^{-9}$).

We compared Model 3 with Model 1 in predicting depression scores and found no difference ($\chi^2 = 1.89$, $p = 0.39$). Model 4 was a significantly better predictor of depression scores than Model 3 ($\chi^2 = 37.42$, $p = 7.45 \times 10^{-9}$).

Examining clinical stages as dummy variables we found statistical differences between all stages for depression scores (see Table 8-4 and Figure 8-5). The depression scores rose by 1.34 points from stage 2a to stage 4.

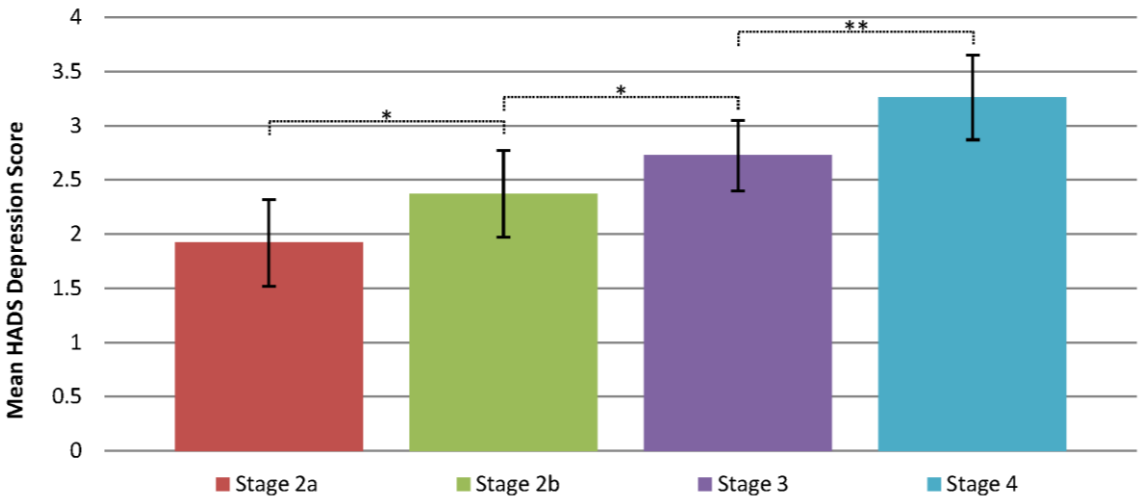


Figure 8-5. Mean HADS depression scores by clinical stage, showing p-values as asterisks.

** $p < 0.01$

* $p < 0.05$

ALS stage	Depression Scores			*Change* in Depression scores when moving to ALS stage								
				Stage 2b			Stage 3			Stage 4		
	Mean	95% CI		Mean	95% CI		Mean	95% CI		Mean	95% CI	
2a	1·92	1·52	2·32	0·45**	0·07	0·83	0·81**	0·41	1·20	1·34**	0·90	1·77
2b	2·37	2·03	2·71				0·36**	0·04	0·67	0·88**	0·51	1·27
3	2·73	2·40	3·05							0·53**	0·18	0·88
4	3·26	2·87	3·65									

Table 8-4. Comparing differences in HADS depression scores across ALS clinical stage using MLE regression

** $p < 0.01$

* $p < 0.05$

8.3.6 Hospital Anxiety and Depression Scale – Anxiety

For anxiety scores, the log likelihood for *Null* Model 1 was -2126.53 and for Model 2 was -2123.71. Model 2 was not a better predictor of HADS anxiety scores ($\chi^2 = 0.14$, $p = 0.06$), although the significance test was very close to the alpha level. We compared Model 3 against Model 1 in predicting HADS anxiety scores and found no difference ($\chi^2 = 4.20$, $p = 0.12$), and also found no difference between Model 4 and Model 3 ($\chi^2 = 0.15$, $p = 0.93$) (see Table 8-5).

Examining clinical stages as dummy variables we found statistical differences between stage 2b and 4, and 3 and 4, for anxiety scores. The increase in anxiety scores between stage 2a and stage 4 was 0.43, and did not linearly increase across stages (Figure 8-6).

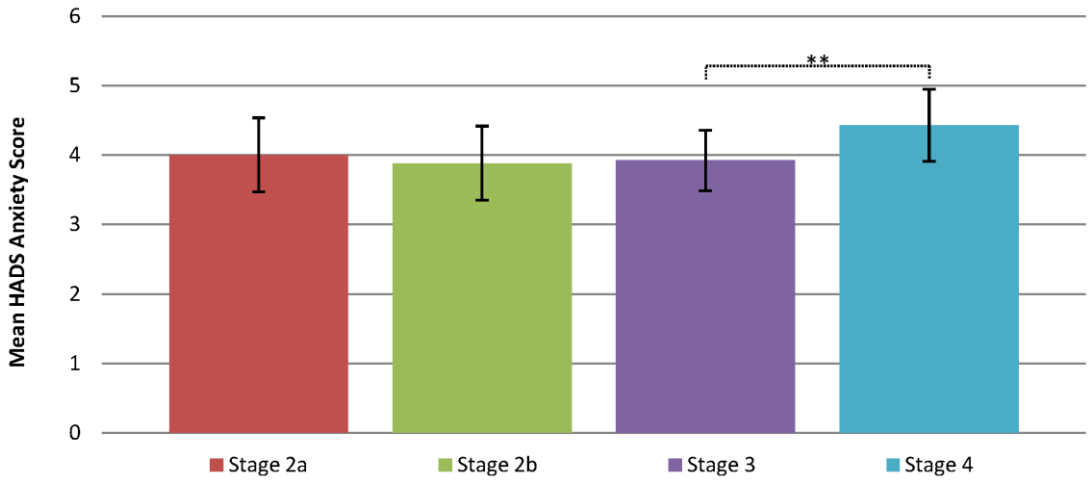


Figure 8-6. Mean HADS anxiety scores by clinical stage, showing p-values as asterisks.

** $p < 0.01$

* $p < 0.05$

ALS stage	Anxiety Scores			*Change* in Anxiety scores when moving to ALS stage								
				Stage 2b			Stage 3			Stage 4		
	Mean	95% CI		Mean	95% CI		Mean	95% CI		Mean	95% CI	
2a	4·00	3·47	4·54	-0·13	-0·63	0·38	-0·08	-0·60	0·45	0·43	-0·15	1·00
2b	3·88	3·43	4·33				0·05	-0·37	0·47	0·55*	0·05	1·05
3	3·93	3·49	4·36							0·51*	0·04	0·97
4	4·43	3·91	4·95									

Table 8-5. Comparing differences in HADS depression scores across ALS clinical stage using MLE regression

** $p < 0.01$

* $p < 0.05$

8.4 Discussion

We have shown that ALS clinical stage corresponds in a predictable and linear manner to differences in HRQoL as assessed by EQ-5D visual analogue and utility scores, and depression scores as assessed by HADS, validating the staging system used, and supporting the use of ALS clinical staging in clinical trials and for allocation of health resources and treatments. The utility estimates at each stage can also be used in models to estimate cost-effectiveness of new therapies for ALS in clinical trials. The reduction in EQ-5D utility and VAS scores with increasingly advanced clinical stage corresponds with similar studies using a functional health state scale[368] and a functional score[163]. However, while previous studies have measured HRQoL in ALS[361, 368-372], the variability of scales used and lack of consensus on what constitutes decline make comparisons difficult. The use of clinical stage overcomes these issues as disease progression is defined rather than a functional proxy. Moreover, this is the largest cohort of individuals with ALS to have been longitudinally assessed against different health states using the EQ-5D scale and thus provides robust progression data.

We modelled ALS clinical stage and disease factors as predictors of EQ-5D health utility, finding that clinical stage was the only significant independent predictor. There are aspects of the EQ-5D measure that limit the amount of testable variance caused by ALS disease progression. First, EQ-5D only measures 5 dimensions of health; thus, symptoms that significantly impact wellbeing, and are present in ALS, such as fatigue, cognitive changes, mood, pain and acute exacerbations are not effectively measured by the EQ-5D scale[373, 374]. Second, the rapid progression of ALS coupled with poor therapeutic options explain why age at onset, sex and disease onset site do not independently influence HRQoL, as quality of life is generally reduced. Third there is a floor effect, with a majority of mid- to late-stage patients scoring in the lower brackets of the individual EQ-5D dimensions, something seen in similar studies[361, 368]. Thus, quantifying the HRQoL of ALS patients using the EQ-5D may be limited by the poor sensitivity in the latter stages of the disease.

Analysis of the individual dimensions of EQ-5D show the greater impact of ALS severity on the ‘physical health’ metrics compared to ‘mental health’ metrics. Predictably, ALS clinical stage was much more effective at predicting the Self Care, Mobility and Usual Activities dimensions than the

Pain/Discomfort, and Anxiety/Depression dimensions of EQ-5D. This is consistent with both studies assessing HRQoL in ALS[368], and a plethora of other neuromuscular disorders, such as fascioscapulohumoral dystrophy and myasthenia gravis [361, 375]. Interestingly, this is in contrast to the significant mental health impact of neurodegenerative disorders, such as Parkinson's, or Huntington's disease, and despite the fact that subclinical frontotemporal dysfunction may be present in up to 50% of those with ALS [376, 377]. However, using the ALS-specific HADS tool to measure mental health, we found depression significantly increased with clinical stage. Research examining the effects of ALS disease progression on mood has also provided conflicting evidence for[378] and against[379] an increase in depression and is entangled with frontotemporal dementia (FTD) related apathy[380, 381], the neurodegenerative companion of ALS.

There are a number of limitations to this study. First, there were two factors acting to reduce statistical power: loss of patients through death, and slow progressing patients who did not reach later disease stages. Nevertheless, by using a longitudinal study we have been able to show unique temporal changes in HRQoL that would not be seen in a cross-sectional study. Second, there is potential for confounding bias, as many variables were not accounted for in the model; for example, we did not include comorbidities, depression, fatigue, or chronic pain as influencers of HRQoL. Third, we did not collect clinical stage prospectively but derived it from existing data. Our own analyses suggest a greater than 92% correlation between actual stage and stage estimated from ALSFRS-R, and it is possible some patients may have been misclassified as a result. A small misclassification rate is unlikely to impact the overall findings.

Using data from a large, multicentre clinical trial, we have shown that clinical stage can be used to model the cost-effectiveness of ALS therapies. The study has acted as a further validation of the King's ALS Staging System, showing that EQ-5D and visual analogue scores differ significantly between stages 2a, 2b, 3 and 4, with more advanced disease stages corresponding to poorer quality of life. The study of longitudinal changes in EQ-5D in patients with ALS provides a framework for health economics analysis of ALS clinical trials data.

Chapter 9 Final Discussion

9.1 Summary of findings

The genotype analyses compared allele frequencies between ALS cases and controls, using logistic and linear regression for genotypes, AOO and survival. The major finding from the genotype analyses was a significant genome-wide association of a C9ORF72 haplotype with ALS, after cases with the C9ORF72 hexanucleotide mutations were removed. I labelled this the residual association as it identifies disease-causing genetic variation unexplained by the known hexanucleotide mutation and which arises from a haplotype distinct from the hexanucleotide-mutation-related haplotype.

Using a genome-wide association study approach, certain genes became statistically significant once ALS was stratified by onset location (limb/bulbar/respiratory). One gene that significantly associated with limb onset cases had a significant epistatic relationship with the C9ORF72 hexanucleotide mutation.

The genotype analyses also confirmed the association of a previously identified 15-SNP ATXN2 haplotype, in UK and Netherland ALS populations. In thesis' dataset I was able to find a more statistically significant haplotype consisting of 47 SNPs associated with ALS. Analyses of PLCD1 did not identify a polymorphism or haplotype that associated with the gene, despite increasing evidence that it is involved in ALS (see section 8.3).

Using the bioinformatic tool ACTA I estimated that a significant proportion of the AOO variance (in the international dataset) and survival (in the UK dataset) can be explained by genetics. I found a SNP significantly predicted AOO in the UK using linear regression, after adjusting for multiple-testing. The four most significant SNPs had an interesting epistatic relationship with an intergenic SNP on chromosome 7. Survival and onset location did not reveal a SNP that showed a statistically significant association with ALS after adjusting for multiple-testing, but did reveal interesting epistasis interactions. The gene FAM195A significantly associated with limb-onset and

has been previously implicated in ALS. Furthermore, we found that it has a significant interaction with the C9ORF72 hexanucleotide mutation.

The gene expression analyses compared spinal regions between ALS cases and controls stratified by anatomy, disease spread for upper limb onset cases, and pathological severity. Expression analyses by anatomy identified several known functional pathological pathways in ALS. The most significant of these were glycosylation and transmembrane activity, calcium binding, regulation of muscle contraction, response to external stimulus & apoptosis, oxidoreductase and flavoproteins, and regulation of insulin, peptide and hormone secretion. I also identified symporter activity, glycan metabolic processes, and pleckstrin homology, as important pathological indicators in the gene expression data. Many of these functions have been previously implicated in ALS pathology.

I applied the same methods to examine gene expression by spread and pathological severity. Using histological and historical clinical data I was able to establish that for four upper-limb cases the disease began in the cervical spinal section and moved caudally quicker than rostrally. However, the gene expression analyses failed to show convincing differences in gene expression profiles although one gene (SLC1A7) consistently differentially expressed across the spinal cord regardless of disease stage. The SLC1A7 protein function is as a glutamate transporter and may relate to the excitotoxicity pathology of ALS.

Analysis looking at gene expression across categories of pathological severity did reveal differences in profile. The pathologically affected region defined as mild-moderate exhibited differentially expressing genes that largely fell into one of three clusters; (a) angiogenesis and blood vessel development function, (b) glycoprotein-based membrane activity, and (c) leukocyte response. Examination of these clusters in increasingly severe pathological categories showed that their enrichment score (representing the presence of that functional pathway) diminished over time. On deeper qualitative inspection however it appears that immune responses are more present in comparison to glycoprotein activity.

Analysing patterns of change in gene expression across categories of pathological severity revealed four genes in which their expression statistically changed. It also replicated two genes previously identified in having putative involvement in ALS. One of these genes may have an interesting interactive relationship with ANGPT2, also implicated in the gene expression analyses, which may be important to ALS. Cluster analyses of the gene expression based on their patterns of change revealed several interesting functional categories, the two most interesting perhaps is an up-down change in leukocyte response which follows closely an up-down change in transcription activity, indicating that the two functions may influence one another

Using non-genetic methods, I tested the hypothesis that clinical staging can be a predictor of disease progression. In this instance disease progression was measured using health utility, health-related quality of life, depression and anxiety. We found that staging ALS is more predictive of measuring disease progression in comparison to using absolute time. This means that ALS progresses at predictable stages when time is standardised. However while this is true statistically, there is much variability between patients in terms of disease time that is unaccounted for.

9.2 Genetic and spread

A major objective of this thesis was to examine two major characteristics of the ALS, its focal genetic pathogenesis and its widespread pathological spread, and analyse how they are linked. These are its (A) genetic basis and (B) its seemingly radial spread from a focal onset location. It is very likely that genetics affect spread because there are genetic modifiers that hasten AOO and increase the belligerence of the disease, if the belligerence of the disease can be defined as the speed and extent of spread. Here I would define AOO not as the age at which the disease ignites but rather when it first becomes noticeable. If ALS spread is a key factor of severity and it

is progressive before the disease becomes symptomatic, then this spread should affect AOO as well as survival.

As we see in Chapter 8, the stages of ALS disease progression are relatively predictable within a standardised timeframe. We are also able to conclude that if disease progresses quickly from one onset location to a second, then the disease will be quick throughout its entire course. Alternatively, if it is slow to begin with then it will have a slow course. In this UK dataset examined here I identified several polymorphisms which seem to aggravate the disease towards death. I also reviewed genetic findings which support the evidence of disease modifying genes quickening the time between stages of ALS.

Analysing the effects of genetics on ALS survival time and AOO, I found several interesting epistatic relationships between SNPs and the most prevalent genetic mutation known to ALS, the C9ORF72 hexanucleotide repeat. Complete functional annotation of the C9ORF72 gene is yet incomplete so I am unable to speculate whether or how these interactions are real. But I also found a residual association with ALS at the C9ORF72 locus which predicts greater than 2 repeat lengths for the C9ORF72 hexanucleotide repeat. Either >2 repeats is a part of residual disease-associating haplotype, or there is a separate unexplained pathological mechanism caused by repeat, for example an alternative repeat sequence or bi-allelic homozygosity >2 repeats. It is interesting that an alternative C9ORF72 repeat may be dose-dependent on the number of repeats (previously thought to be non-pathogenic). Taken together with the now known somatic heterogeneity of C9ORF72, it acts as a good example of how a genetic mutations' variability can cause variation in disease spread and perhaps even lead to focal onset by manifesting strong dosage in a particular weak-spot of the spinal cord or motor cortex.

To examine the seeming dualistic discrepancy of a genetic but focal-onset disease I aimed to analyse the interaction between genetics and spread using gene expression analyses. My objective was to identify abnormal gene expression in a spinal region that progressively changed throughout disease spread and severity. This region could be seen as weak-spots, defined as cells and/or regions that are less protected against a specific genetic mutation(s), or possible 'external factors' of toxins or antigens, that facilitate the activation of ALS at a specific location(s) in the spinal cord or motor cortex.

The gene expression analyses identified many genes that abnormally expressed at different pathological stages of the disease, including genes that showed evidence of dose-dependent changes as the disease moved from mild to severe. The genes and functions identified replicated known pathways in ALS disease. Moving from mild to severe pathological categories, the expression profiles changed from very definitive very significant functional categories implicating many genes, towards more ill-defined small gene-number categories. Patterns of gene expression change supported pre-existing hypothesis of how some of these function may have toxic gain-of-function to begin with, which eventually become under-regulated. Furthermore I was able to identify functional overlap between the three major profiles identified; angiogenesis, glycoproteins and immune response, the latter showing a statistically definitive pattern of change as a consequence of increasing pathological severity.

The following is a simplistic example but the findings allow speculation on the manifestation of ALS. Say a person inherits a genetic mutation affecting either angiogenesis or glycoproteins. This elicits an immune response but not one that requires a strong presence in the C.N.S. As age increases immunocompetence declines and the genetic mutations compromises both aberrant blood brain barrier (BBB) signalling and muscle atrophy, causing an increased amount of leukocytes bypass the BBB into the C.N.S. Initially these are beneficial but overtime, and likely influenced by an array genetic variation in a person, the activity of leukocytes become toxic. These cause an array of cytotoxic events eventually killing motor neurons leading to the fast cascading disease of ALS.

The concept of genes has moved away from them being stationary coding units. Analysing the interactions between genetics, RNA, and pathology, may unveil hidden mechanisms of disease causation and manifestation in ALS. We know they play a definitive role in the onset of ALS and it is likely they will influence its spread in terms of time and severity.

9.3 Genetic candidates

Genes C9ORF72 and SLC1A7 have already been discussed as gene candidates above. PLCD1 is a strong candidate for ALS despite not finding a haplotype associated with the disease in this dataset. Papers show its association with ALS and that its ablation worsens SOD1-related ALS phenotypes. In the gene expression analysis genes of the same family showed significant differential expression especially mild to moderate stages of the disease. These were PLCD3, PLCD4, and PLCXD3.

CTBP2 may modulate ALS survival; it is known to suppress transcription exposing a possible route of pathogenesis. FAM19A5 and LOC284993 were found to be interesting as their association with ALS significantly increased when only limb-onset patients were examined, and was statistically related to the C9ORF72 repeat mutation. These genes are functionally uncharacterised.

Other genes from the gene expression analysis were MSRA which showed a network relationship with KIF3A, and PLCD3 showing a network relationship with ALS gene ITPR1. MSRA is involved in suloxide reductase whereas PLCD3 is likely to be involved in responses to extracellular stimuli. The gene ADA showed a functional relationship with the ALS2 gene. ADA contributes to cellular signalling responses, which are somewhat already implicated in ALS pathology. UFD1L also showed a showed a protein interaction relationship with ALS VCP. The CYB5R1 also differentially expresses and interacts with ubiquitin proteins, which may group UFD1L, VCP, CYB5R1 and UBQLN under the same ALS pathological pathway.

Gene expression by pathology revealed several interesting candidates. Genes TMEM132E, HAMP, ZKSAN1 and REM1 all showed changes in expression as the disease progressed. Genes SUSD1, CRIM1, KDR, and RAMP3 have all been previously implicated in ALS research and I found them to differentially expression when just taking hold of a spinal region. GEMIN5, involved in pre-mRNA splicing, is a good candidate as interacts with SMN1 and was found to differentially express in moderately effected spinal regions.

ANGPT2 has a very close relationship with ALS gene VEGFA, in that its immune-based activity depends on VEGFA activity. This was most salient at the most severest stage of the disease, suggesting ANGPT2's candidacy as survival modifier.

9.4 Research obstacles and future direction

Using genotypes from blood samples does not allow anatomical specificity. Therefore it does not allow me to determine which genetic mutations or risk variants are important where. We can infer this using expression data and location but the activity of these regions are not often well categorised. For large effect size mutations, such as the C9ORF72 hexanucleotide mutation, it would be useful for studies to examine DNA from multiple anatomical regions involved in ALS. This is important for this thesis, as I am unable to confidently conclude whether the residual association at C9ORF72 was not the consequence of the somatic heterogeneity of the mutation. Although, this is unlikely as the residual association is a distinct haplotype background to that of the mutation.

I identified that a significant proportion of survival variance in ALS can be accounted for by genetics. I did not find the same for AOO. It may be that survival in ALS is always modulated by genetic variance, whereas AOO can be. In a large population AOO genetic variance as whole did not show significant association with, but we are able identify specific nucleotide polymorphisms in certain patients. Further annotation of the genes associated with AOO (for example the statistical interaction between FAM195A and C9ORF72) are required to construct a hypothesis on why and how this relationship is important in ALS.

The extent to which genotype analyses will have utility in complex neurodegenerative diseases is becoming increasingly questioned. I believe using haplotype association testing in relation to ATXN2 and PLCD1 was a good method to substantiate their association with ALS, primarily because their underlying effects are believed to genetic, in the form a nucleotide repeat. Repeats can arise from haplotypic inheritance, which I aimed to identify here. But this method relies on the

premise that common variants identify underlying somewhat complex haplotypes, which may not be true. To counteract this, a method utilising rare variants allowing for haplotype variation across populations may have utility when dealing with mutations of low penetrance.

The problem of genetic variance hidden in genotype analyses was somewhat tackled using the RNA-seq dataset, in an attempt to identify variation within known ALS genes that could help in future research, where common variant genotyping has been unsuccessful. ELP3 for example shows complex transcript variation, which has been a focus of our collaborators. Knowing that ELP3 association is unlikely to be due to a significant DNA mutation and gene expression analyses do not show it to be abnormally expressed but that protein expression of ELP3 is significantly reduced in ALS patients, warrants further examination into this transcript variation.

The Illumina BeadChip platform examines transcript variation to some extent but is not as detailed as RNA-seq. This in itself is a disadvantage of using gene expression to examine genetic mutations. Gene expression informs us on which genes are abnormally activated (or deactivated) in relation to ALS but does not inform us on whether a genetic abnormality is due to gene or transcript mutation, or whether it is responding to a pathological environment. The combination of contextually using gene expression to identify gene candidates and then RNA-seq to deepen the perspective on those candidates may be a beneficial approach in the future.

In ALS there is an added problem in using gene expression, as RNA from motor neurons, when the disease is at its most severe, should be almost not existent. It is questionable how informative gene expression analysis is at these stages. The most informative pathological stage in my analysis was when the disease was just beginning to take hold of a spinal region. Future gene expression research should take this into account when simply comparing across spinal segments. Bringing in eQTL data across spinal cord levels would also be informative and may clarify some of the SNPs which are associated with ALS.

The clinical staging of ALS is not informed by genetic information. This is currently being developed. It would be interesting to include genes that we know create severe or mild ALS phenotypes and analyse how these change the time a patient stays in each stage. Onset-location is currently included in the clinical staging model, differentiating between limb and bulbar onset.

However it could be informative of how ALS spreads, if we analysed staging times by imposing on the model an onset location and perhaps clinical symptomatological progression. A direction in which clinical staging could be taken therefore is to stratify populations by genetic mutation, onset-location and spread, which may allow a greater level of specificity in predicting a patient's course of disease.

9.5 Final summary

This thesis examined the genetics and spread of ALS. Its findings supported the association of individual genes implicated in the disease, and found no evidence for others. I found that a significant proportion of survival in ALS can be accounted for by genetics, and identified SNPs which predict AOO and survival. I identified a residual association at C9ORF72 that is not due to the known hexanucleotide repeat mutation, and that this residual association also predicts greater than two hexanucleotide repeats. Furthermore, we identified a gene that shows statistical epistasis with the known C9ORF72 hexanucleotide repeat. A gene expression analysis by spinal cord anatomy, spread and pathology, identified profiles of genes strongly implicating blood vessel development and angiogenesis, glycoprotein activity, and leukocyte migration. Some gene expression profiles were consistent throughout the course of the disease, while others changed in agreement with predictions of how cell function may alter as a consequence of ALS involvement. This thesis also examined disease progression, a putative indicator of spread, with the King's ALS Clinical Staging system, finding that it predicted changes in health utility, motor function and psychometric function, over time.

The overall finding of thesis is that spread in ALS predictably affects not only physiology but gene expression, and that this likely to be modulated by genetic mutations.

References

1. Andersen PM, Al-Chalabi A. Clinical genetics of amyotrophic lateral sclerosis: what do we really know? *Nat Rev Neurol* 2011; **7**(11): 603-15.

2. van Blitterswijk M, Landers J. RNA processing pathways in amyotrophic lateral sclerosis. *neurogenetics* 2010; **11**(3): 275-90.

3. Deng HX, Chen W, Hong ST, et al. Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature* 2011; **477**(7363): 211-5.

4. van Blitterswijk M, van Es MA, Hennekam EAM, et al. Evidence for an oligogenic basis of amyotrophic lateral sclerosis. *Human Molecular Genetics* 2012; **21**(17): 3776-84.

5. Kurland LT, Mulder DW. Epidemiologic Investigations of Amyotrophic Lateral Sclerosis: 2. Familial Aggregations Indicative of Dominant Inheritance Part I. *Neurology* 1955; **5**(3): 182-96.

6. Hentati A, Bejaoui K, Pericak-Vance MA, et al. Linkage of recessive familial amyotrophic lateral sclerosis to chromosome 2q33-q35. *Nat Genet* 1994; **7**(3): 425-8.

7. Simpson CL, Al-Chalabi A. Amyotrophic lateral sclerosis as a complex genetic disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 2006; **1762**(11-12): 973-85.

8. Fang F. Epidemiologic studies of amytrophic lateral sclerosis. 2010.

9. Johnston C, Stanton B, Turner M, et al. Amyotrophic lateral sclerosis in an urban setting: a population based study of inner city London. *J Neurol* 2006; **253**: 1642 - 3.

10. Pasinelli P, Brown RH. Molecular biology of amyotrophic lateral sclerosis: insights from genetics. *Nat Rev Neurosci* 2006; **7**(9): 710-23.

11. Mulder DW, Kurland LT, Offord KP, Beard CM. Familial adult motor neuron disease: amyotrophic lateral sclerosis. *Neurology* 1986; **36**(4): 511-7.

12. Robberecht WMDP, Aguirre TM, Van Den Bosch LP, Tilkin PR, Cassiman JJMDP, Matthijs GP. D90A heterozygosity in the SOD1 gene is associated with familial and apparently sporadic amyotrophic lateral sclerosis. *Neurology* 1996; **47**(5): 1336-9.

13. Al-Chalabi A, Fang F, Hanby MF, et al. An estimate of amyotrophic lateral sclerosis heritability using twin data. *Journal of Neurology, Neurosurgery & Psychiatry* 2010; **81**: 1324-6.

14. Wroe R, Wai-Ling Butler A, Andersen PM, Powell JF, Al-Chalabi A. ALSOD: The Amyotrophic Lateral Sclerosis Online Database. *Amyotrophic Lateral Sclerosis* 2008; **9**(4): 249-50.

15. Aran F. Recherches sur une maladie non encore decrite du systeme musculaire (atrophie musculaire progressive). *Arch Gen Med* 1850; **14**: 5 - 35.

16. Siddique T, Figlewigz DA, Pericak-Vance MA, et al. Linkage of a Gene Causing Familial Amyotrophic Lateral Sclerosis to Chromosome 21 and Evidence of Genetic-Locus Heterogeneity. *New England Journal of Medicine* 1991; **324**(20): 1381-4.

17. Rosen DR, Siddique T, Patterson D, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 1993; **362**(6415): 59-62.

18. Turner BJ, Talbot K. Transgenics, toxicity and therapeutics in rodent models of mutant SOD1-mediated familial ALS. *Progress in Neurobiology* 2008; **85**(1): 94-134.

19. Shaw CE, Enayat ZE, Chioza BA, et al. Mutations in all five exons of SOD-1 may cause ALS. *Annals of Neurology* 1998; **43**(3): 390-4.

20. Valdmanis PNB, Rouleau GAMDP. Genetics of familial amyotrophic lateral sclerosis. *Neurology* 2008; **70**(2): 144-52.

21. Cudkowicz ME, McKenna-Yasek D, Sapp PE, et al. Epidemiology of mutations in superoxide dismutase in amyotrophic lateal sclerosis. *Annals of Neurology* 1997; **41**(2): 210-21.

22. Radunovíc A, Leigh PN. Cu/Zn superoxide dismutase gene mutations in amyotrophic lateral sclerosis: correlation between genotype and clinical features. *Journal of Neurology, Neurosurgery & Psychiatry* 1996; **61**(6): 565-72.

23. Juneja T, Pericak-Vance MA, Laing NG, Dave S, Siddique T. Prognosis in Familial Amyotrophic Lateral Sclerosis. *Neurology* 1997; **48**(1): 55-7.

24. Broom WJ, Johnson DV, Auwarter KE, et al. SOD1A4V-mediated ALS: Absence of a closely linked modifier gene and origination in Asia. *Neuroscience Letters* 2008; **430**(3): 241-5.

25. DiDonato M, Craig L, Huff ME, et al. ALS Mutants of Human Superoxide Dismutase Form Fibrous Aggregates Via Framework Destabilization. *Journal of Molecular Biology* 2003; **332**(3): 601-15.

26. Andersen PM, Forsgren L, Binzer M, et al. Autosomal recessive adult-onset amyotrophic lateral sclerosis associated with homozygosity for Asp90Ala CuZn-superoxide dismutase mutation. *Brain* 1996; **119**(4): 1153-72.

27. Al-Chalabi A, Andersen PM, Chioza B, et al. Recessive Amyotrophic Lateral Sclerosis Families with the D90A SOD1 Mutation Share a Common Founder: Evidence for a Linked Protective Factor. *Human Molecular Genetics* 1998; **7**(13): 2045-50.

28. Själander A, Beckman G, Deng H-X, lqbal Z, Tainer JA, Siddique T. The D90A mutation results in a polymorphism of Cu, Zn superoxide dismutase that is prevalent in northern Sweden and Finland. *Human Molecular Genetics* 1995; **4**(6): 1105-8.

29. Gurney M, Pu H, Chiu A, et al. Motor neuron degeneration in mice that express a human Cu,Zn superoxide dismutase mutation. *Science* 1994; **264**(5166): 1772-5.

30. Durham HD, Roy J, Dong L, Figlewicz DA. Aggregation of Mutant Cu/Zn Superoxide Dismutase Proteins in a Culture Model of ALS. *Journal of Neuropathology & Experimental Neurology* 1997; **56**(5): 523-30.

31. Higgins CMJ, Jung C, Ding H, Xu Z. Mutant Cu, Zn Superoxide Dismutase that Causes Motoneuron Degeneration Is Present in Mitochondria in the CNS. *J Neurosci* 2002; **22**(6): 215RC-
.

32. Yim MB, Kang JH, Yim HS, Kwak HS, Chock PB, Stadtman ER. A gain-of-function of an amyotrophic lateral sclerosis-associated Cu,Zn-superoxide dismutase mutant: An enhancement of free radical formation due to a decrease in Km for hydrogen peroxide. *Proceedings of the National Academy of Sciences of the United States of America* 1996; **93**(12): 5709-14.

33. Beckman JS, Estévez AG, Crow JP, Barbeito L. Superoxide dismutase and the death of motoneurons in ALS. *Trends in Neurosciences* 2001; **24**(11): S15-S20.

34. Greenway MJM, Alexander MDM, Ennis SP, et al. A novel candidate region for ALS on chromosome 14q11.2. *Neurology* 2004; **63**(10): 1936-8.

35. Hayward CP, Colville SB, Swingler RJMD, Brock DJHP. Molecular genetic analysis of the APEX nuclease gene in amyotrophic lateral sclerosis. *Neurology* 1999; **52**(9): 1899-901.

36. Gellera C, Colombrita C, Ticozzi N, et al. Identification of new ANG gene mutations in a large cohort of Italian patients with amyotrophic lateral sclerosis. *neurogenetics* 2008; **9**(1): 33-40.

37. Fernández-Santiago R, Hoenig S, Lichtner P, et al. Identification of novel Angiogenin gene missense variants in German patients with amyotrophic lateral sclerosis. *Journal of Neurology* 2009; **256**(8): 1337-42.

38. Paubel A, Violette J, Amy M, et al. Mutations of the ANG Gene in French Patients With Sporadic Amyotrophic Lateral Sclerosis. *Arch Neurol* 2008; **65**(10): 1333-6.

39. McLaughlin RL, Phukan J, McCormack W, et al. Angiogenin Levels and ANG Genotypes: Dysregulation in Amyotrophic Lateral Sclerosis. *PLoS ONE* 2010; **5**(11): e15402.

40. Greenway M, Andersen P, Russ C, et al. ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat Genet* 2006; **38**: 411 - 3.

41. Greenway M, Alexander M, Ennis S, et al. A novel candidate region for ALS on chromosome 14q11.2. *Neurology* 2004; **63**: 1936 - 8.

42. Leigh P, Whitwell H, Garofalo O, et al. Ubiquitin-immunoreactive intraneuronal inclusions in amyotrophic lateral sclerosis. Morphology, distribution, and specificity. *Brain* 1991; **114**(Pt 2): 775 - 88.

43. Lansbury PT, Lashuel HA. A century-old debate on protein aggregation and neurodegeneration enters the clinic. *Nature* 2006; **443**(7113): 774-9.

44. Sreedharan J, Blair I, Tripathi V, et al. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* 2008; **319**: 1668 - 72.

45. Gitcho MA, Baloh RH, Chakraverty S, et al. TDP-43 A315T mutation in familial motor neuron disease. *Annals of Neurology* 2008; **63**(4): 535-8.

46. Ou S, Wu F, Harrich D, Garcia-Martinez L, Gaynor R. Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. *J Virol* 1995; **69**(6): 3584-96.

47. Wang H-Y, Wang IF, Bose J, Shen CKJ. Structural diversity and functional implications of the eukaryotic TDP gene family. *Genomics* 2004; **83**(1): 130-9.

48. Buratti E, Dork T, Zuccato E, Pagani F, Romano M, Baralle FE. Nuclear factor TDP-43 and SR proteins promote in vitro and in vivo CFTR exon 9 skipping. *EMBO J* 2001; **20**(7): 1774-84.

49. Mackenzie I, Bigio E, Ince P, et al. Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. *Ann Neurol* 2007; **61**: 427 - 34.

50. Benajiba L, Le Ber I, Camuzat A, et al. TARDBP mutations in motoneuron disease with frontotemporal lobar degeneration. *Annals of Neurology* 2009; **65**(4): 470-3.

51. Buratti E, Brindisi A, Giombi M, Tisminetzky S, Ayala YM, Baralle FE. TDP-43 Binds Heterogeneous Nuclear Ribonucleoprotein A/B through Its C-terminal Tail. *Journal of Biological Chemistry* 2005; **280**(45): 37572-84.

52. Huang R, Fang D-F, Ma M-Y, et al. TARDBP gene mutations among Chinese patients with sporadic amyotrophic lateral sclerosis. *Neurobiology of Aging*; **In Press, Corrected Proof**.

53. Lagier-Tourenne C, Cleveland DW. Rethinking ALS: The FUS about TDP-43. *Cell* 2009; **136**(6): 1001-4.

54. Kabashi E, Valdmanis P, Dion P, et al. TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat Genet* 2008; **40**: 572 - 4.

55. Daoud H, Valdmanis PN, Kabashi E, et al. Contribution of TARDBP mutations to sporadic amyotrophic lateral sclerosis. *Journal of Medical Genetics* 2009; **46**(2): 112-4.

56. Neumann M, Sampathu D, Kwong L, et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* 2006; **314**: 130 - 3.

57. Van Deerlin VM, Leverenz JB, Bekris LM, et al. TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *The Lancet Neurology* 2008; **7**(5): 409-16.

58. Ruddy DM, Parton MJ, Al-Chalabi A, et al. Two families with familial amyotrophic lateral sclerosis are linked to a novel locus on chromosome 16q. *American journal of human genetics* 2003; **73**(2): 390-6.

59. Vance C, Rogelj B, Hortobagyi T, et al. Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6. *Science* 2009; **323**(5918): 1208-11.

60. Kwiatkowski TJ, Jr., Bosco DA, LeClerc AL, et al. Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. *Science* 2009; **323**(5918): 1205-8.

61. Mackenzie IRA, Rademakers R, Neumann M. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *The Lancet Neurology* 2010; **9**(10): 995-1007.

62. Hewitt C, Kirby J, Highley JR, et al. Novel FUS/TLS Mutations and Pathology in Familial and Sporadic Amyotrophic Lateral Sclerosis. *Arch Neurol* 2010; **67**(4): 455-61.

63. Pasinelli P, Brown R. Molecular biology of amyotrophic lateral sclerosis: insights from genetics. *Nat Rev Neurosci* 2006; **7**: 710 - 23.

64. Lagier-Tourenne C, Polymenidou M, Cleveland DW. TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Human Molecular Genetics* 2010; **19**(R1): R46-R64.

65. Munoz D, Neumann M, Kusaka H, et al. FUS pathology in basophilic inclusion body disease. *Acta Neuropathologica* 2009; **118**(5): 617-27.

66. Neumann M, Rademakers R, Roeber S, Baker M, Kretzschmar HA, Mackenzie IRA. A new subtype of frontotemporal lobar degeneration with FUS pathology. *Brain* 2009; **132**(11): 2922-31.

67. Yang L, Embree LJ, Tsai S, Hickstein DD. Oncoprotein TLS Interacts with Serine-Arginine Proteins Involved in RNA Splicing. *Journal of Biological Chemistry* 1998; **273**(43): 27761-4.

68. Hicks GG, Singh N, Nashabi A, et al. Fus deficiency in mice results in defective B-lymphocyte development and activation, high levels of chromosomal instability and perinatal death. *Nat Genet* 2000; **24**(2): 175-9.

69. Uranishi H, Tetsuka T, Yamashita M, et al. Involvement of the Pro-oncoprotein TLS (Translocated in Liposarcoma) in Nuclear Factor-κB p65-mediated Transcription as a Coactivator. *Journal of Biological Chemistry* 2001; **276**(16): 13395-401.

70. Corrado L, Del Bo R, Castellotti B, et al. Mutations of FUS gene in sporadic amyotrophic lateral sclerosis. *Journal of Medical Genetics* 2010; **47**(3): 190-4.

71. Yang S, Warraich ST, Nicholson GA, Blair IP. Fused in sarcoma/translocated in liposarcoma: A multifunctional DNA/RNA binding protein. *The International Journal of Biochemistry & Cell Biology* 2010; **42**(9): 1408-11.

72. Millecamps S, Salachas F, Cazeneuve C, et al. SOD1, ANG, VAPB, TARDBP, and FUS mutations in familial amyotrophic lateral sclerosis: genotype–phenotype correlations. *Journal of Medical Genetics* 2010; **47**(8): 554-60.

73. Maruyama H, Morino H, Ito H, et al. Mutations of optineurin in amyotrophic lateral sclerosis. *Nature* 2010; **465**(7295): 223-6.

74. Mitchell J, Paul P, Chen H-J, et al. Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase. *Proceedings of the National Academy of Sciences* 2010; **107**(16): 7556-61.

75. Watts GDJ, Wymer J, Kovach MJ, et al. Inclusion body myopathy associated with Paget disease of bone and frontotemporal dementia is caused by mutant valosin-containing protein. *Nat Genet* 2004; **36**(4): 377-81.

76. Johnson JO, Mandrioli J, Benatar M, et al. Exome Sequencing Reveals VCP Mutations as a Cause of Familial ALS. *Neuron* 2010; **68**(5): 857-64.

77. Morita M, Al-Chalabi A, Andersen PM, et al. A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia. *Neurology* 2006; **66**(6): 839-44.

78. Vance C, Al-Chalabi A, Ruddy D, et al. Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2–21.3. *Brain* 2006; **129**(4): 868-76.

79. van Es MA, Veldink JH, Saris CGJ, et al. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat Genet* 2009; **41**(10): 1083-7.

80. Shatunov A, Mok K, Newhouse S, et al. Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *The Lancet Neurology* 2010; **9**(10): 986-94.

81. Laaksovirta H, Peuralinna T, Schymick JC, et al. Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. *The Lancet Neurology* 2010; **9**(10): 978-85.

82. DeJesus-Hernandez M, Mackenzie Ian R, Boeve Bradley F, et al. Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron* 2011; **72**(2): 245-56.

83. Renton Alan E, Majounie E, Waite A, et al. A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron* 2011; **72**(2): 257-68.

84. Majounie E, Renton AE, Mok K, et al. Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *The Lancet Neurology* 2012; **11**(4): 323-30.

85. Okado-Matsumoto A, Fridovich I. Amyotrophic lateral sclerosis: A proposed mechanism. *Proceedings of the National Academy of Sciences* 2002; **99**(13): 9010-4.

86. Nishimura A, Mitne-Neto M, Silva H, et al. A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *Am J Hum Genet* 2004; **75**: 822 - 31.

87. Hadano S, Hand C, Osuga H, et al. A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nat Genet* 2001; **29**: 166 - 73.

88. Yang Y, Hentati A, Deng H, et al. The gene encoding alsin, a protein with three guanine-nucleotide exchange factor domains, is mutated in a form of recessive amyotrophic lateral sclerosis. *Nat Genet* 2001; **29**: 160 - 5.

89. Chen Y, Bennett C, Huynh H, et al. DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *Am J Hum Genet* 2004; **74**: 1128 - 35.

90. Figlewicz DA, Krizus A, Martinoli MG, et al. Variants of the heavy neurofilament subunit are associated with the development of amyotrophic lateral sclerosis. *Human Molecular Genetics* 1994; **3**(10): 1757-61.

91. Al-Chalabi A, Andersen PM, Nilsson P, et al. Deletions of the Heavy Neurofilament Subunit Tail in Amyotrophic Lateral Sclerosis. *Human Molecular Genetics* 1999; **8**(2): 157-64.

92. Tomkins J, Usher P, Slade J, et al. Novel insertion in the KSP region of the neurofilament heavy gene in amyotrophic lateral sclerosis (ALS). *Neuroreport* 1998; **9**: 3967 - 70.

93. Vechio JD, Bruijn LI, Xu Z, Brown RH, Cleveland DW. Sequence variants in human neurofilament proteins: Absence of linkage to familial amyotrophic lateral sclerosis. *Annals of Neurology* 1996; **40**(4): 603-10.

94. Rooke K, Figlewicz DA, Han F-y, Rouleau GA. Analysis of the KSP repeat of the neurofilament heavy subunit in familial amyotrophic lateral sclerosis. *Neurology* 1996; **46**(3): 789-90.

95. Garcia ML, Singleton AB, Hernandez D, et al. Mutations in neurofilament genes are not a significant primary cause of non-SOD1-mediated amyotrophic lateral sclerosis. *Neurobiology of Disease* 2006; **21**(1): 102-9.

96. Wong NKYM, He BEIPMDP, Strong MJM. Characterization of Neuronal Intermediate Filament Protein Expression in Cervical Spinal Motor Neurons in Sporadic Amyotrophic Lateral Sclerosis (ALS). *Journal of Neuropathology & Experimental Neurology* 2000; **59**(11): 972-82.

97. Elden AC, Kim H-J, Hart MP, et al. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature* 2010; **466**(7310): 1069-75.

98. Nanetti L, Fancellu R, Tomasello C, Gellera C, Pareyson D, Mariotti C. Rare association of motor neuron disease and spinocerebellar ataxia type 2 (SCA2): a new case and review of the literature. *Journal of Neurology* 2009; **256**(11): 1926-8.

99. Lefebvre S, Bürglen L, Reboullet S, et al. Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* 1995; **80**(1): 155-65.

100. Parsons DW, McAndrew PE, Iannaccone ST, Mendell JR, Burghes AHM, Prior TW. Intragenic telSMN Mutations: Frequency, Distribution, Evidence of a Founder Effect, and Modification of the Spinal Muscular Atrophy Phenotype by cenSMN Copy Number. *The American Journal of Human Genetics* 1998; **63**(6): 1712-23.

101. Corcia P, Mayeux-Portas V, Khoris J, et al. Abnormal SMN1 gene copy number is a susceptibility factor for amyotrophic lateral sclerosis. *Annals of Neurology* 2002; **51**(2): 243-6.

102. Corcia P, Camu W, Halimi JM, et al. SMN1 gene, but not SMN2, is a risk factor for sporadic ALS. *Neurology* 2006; **67**(7): 1147-50.

103. Blauw HM, Al-Chalabi A, Andersen PM, et al. A large genome scan for rare CNVs in amyotrophic lateral sclerosis. *Human Molecular Genetics* 2010; **19**(20): 4091-9.

104. Cronin S, Blauw HM, Veldink JH, et al. Analysis of genome-wide copy number variation in Irish and Dutch ALS populations. *Human Molecular Genetics* 2008; **17**(21): 3392-8.

105. Wain LV, Pedroso I, Landers JE, et al. The Role of Copy Number Variation in Susceptibility to Amyotrophic Lateral Sclerosis: Genome-Wide Association Study and Comparison with Published Loci. *PLoS ONE* 2009; **4**(12): e8175.

106. Oosthuyse B, Moons L, Storkebaum E, et al. Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. *Nat Genet* 2001; **28**(2): 131-8.

107. Subramanian V, Crabtree B, Acharya KR. Human angiogenin is a neuroprotective factor and amyotrophic lateral sclerosis associated angiogenin variants affect neurite

extension/pathfinding and survival of motor neurons. *Human Molecular Genetics* 2008; **17**(1): 130-49.

108. Kieran D, Sebastia J, Greenway MJ, et al. Control of Motoneuron Survival by Angiogenin. *J Neurosci* 2008; **28**(52): 14056-61.

109. Lambrechts D, Storkebaum E, Morimoto M, et al. VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. *Nat Genet* 2003; **34**: 383 - 94.

110. Van Vught PWJ, Sutedja NA, Veldink JH, et al. Lack of association between VEGF polymorphisms and ALS in a Dutch population. *Neurology* 2005; **65**(10): 1643-5.

111. Chen WM, Saeed MM, Mao HMM, et al. Lack of association of VEGF promoter polymorphisms with sporadic ALS. *Neurology* 2006; **67**(3): 508-10.

112. Fernández-Santiago R, Sharma M, Mueller JC, et al. Possible gender-dependent association of vascular endothelial growth factor (VEGF) gene and ALS. *Neurology* 2006; **66**(12): 1929-31.

113. Lambrechts D, Poesen K, Fernández-Santiago R, et al. Meta-analysis of vascular endothelial growth factor variations in amyotrophic lateral sclerosis: increased susceptibility in male carriers of the −2578AA genotype. *Journal of Medical Genetics* 2009; **46**(12): 840-6.

114. Cronin S, Berger S, Ding J, et al. A genome-wide association study of sporadic ALS in a homogenous Irish population. *Human Molecular Genetics* 2008; **17**(5): 768-74.

115. van Es MA, Van Vught PW, Blauw HM, et al. ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *The Lancet Neurology* 2007; **6**(10): 869-77.

116. van Es MA, van Vught PWJ, Blauw HM, et al. Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat Genet* 2008; **40**(1): 29-31.

117. Cronin S, Tomik B, Bradley DG, Slowik A, Hardiman O. Screening for replication of genome-wide SNP associations in sporadic ALS. *Eur J Hum Genet* 2008; **17**(2): 213-8.

118. Bo RD, Ghezzi S, Corti S, et al. DPP6 gene variability confers increased risk of developing sporadic amyotrophic lateral sclerosis in Italian patients. *Journal of Neurology, Neurosurgery & Psychiatry* 2008; **79**(9): 1085.

119. Zhong H, Prentice RL. Correcting “winner's curse” in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epidemiology* 2010; **34**(1): 78-91.

120. Fogh I, D'Alfonso S, Gellera C, et al. No association of DPP6 with amyotrophic lateral sclerosis in an Italian population. *Neurobiology of Aging*; **In Press, Corrected Proof**.

121. Chiò A, Schymick JC, Restagno G, et al. A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. *Human Molecular Genetics* 2009; **18**(8): 1524-32.

122. Xiao-guang L, Jiang-hu Z, Man-qing X, et al. Association between DPP6 polymorphism and the risk of sporadic amyotrophic lateral sclerosis in Chinese patients. *Chinese Medical Journal* 2009; **122**(24): 2989 - 92.

123. Fernández-Santiago R, Sharma M, Berg D, et al. No evidence of association of FLJ10986 and ITPR2 with ALS in a large German cohort. *Neurobiology of Aging*; **In Press, Corrected Proof**.

124. Varoqueaux F, Sons MS, Plomp JJ, Brose N. Aberrant Morphology and Residual Transmitter Release at the Munc13-Deficient Mouse Neuromuscular Synapse. *Mol Cell Biol* 2005; **25**(14): 5973-84.

125. Hawkes NA, Otero G, Winkler GS, et al. Purification and Characterization of the Human Elongator Complex. *Journal of Biological Chemistry* 2002; **277**(4): 3047-52.

126. Winkler GS, Petrakis TG, Ethelberg S, et al. RNA Polymerase II Elongator Holoenzyme Is Composed of Two Discrete Subcomplexes. *Journal of Biological Chemistry* 2001; **276**(35): 32743-9.

127. Winkler GS, Kristjuhan A, Erdjument-Bromage H, Tempst P, Svejstrup JQ. Elongator is a histone H3 and H4 acetyltransferase important for normal histone acetylation levels in vivo. *Proceedings of the National Academy of Sciences of the United States of America* 2002; **99**(6): 3517-22.

128. Huang B, Johansson MJO, Bystrom AS. An early step in wobble uridine tRNA modification requires the Elongator complex. *RNA* 2005; **11**(4): 424-36.

129. Simpson CL, Lemmens R, Miskiewicz K, et al. Variants of the elongator protein 3 (ELP3) gene are associated with motor neuron degeneration. *Human Molecular Genetics* 2009; **18**(3): 472-81.

130. Anderson SL, Coli R, Daly IW, et al. Familial Dysautonomia Is Caused by Mutations of the IKAP Gene. *American journal of human genetics* 2001; **68**(3): 753-8.

131. Barton D, Braet F, Marc J, Overall R, Gardiner J. ELP3 localises to mitochondria and actin-rich domains at edges of HeLa cells. *Neuroscience Letters* 2009; **455**(1): 60-4.

132. Fogh I, Rijsdijk F, Andersen P, et al. Age at onset in sod1-mediated amyotrophic lateral sclerosis shows familiality. *Neurogenetics* 2007; **8**(3): 235-6.

133. Moulard B, Sefiani A, Laamri A, Malafosse A, Camu W. Apolipoprotein E genotyping in sporadic amyotrophic lateral sclerosis: evidence for a major influence on the clinical presentation and prognosis. *Journal of the neurological sciences* 1996; **139**(Supplement 1): 34-7.

134. Zetterberg H, Jacobsson J, Rosengren L, Blennow K, Andersen PM. Association of APOE with age at onset of sporadic amyotrophic lateral sclerosis. *Journal of the neurological sciences* 2008; **273**(1-2): 67-9.

135. Gros-Louis F, Andersen PM, Dupre N, et al. Chromogranin B P413L variant as risk factor and modifier of disease onset for amyotrophic lateral sclerosis. *Proceedings of the National Academy of Sciences* 2009; **106**(51): 21777-82.

136. Mitchell J, Morris A, de Belleroche J. Thioredoxin reductase 1 haplotypes modify familial amyotrophic lateral sclerosis onset. *Free Radical Biology and Medicine* 2009; **46**(2): 202-11.

137. Consortium A. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. *Neurobiology of Aging* 2013; **34**(1): 357.e7-.e19.

138. Landers JE, Melki J, Meininger V, et al. Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proceedings of the National Academy of Sciences* 2009; **106**(22): 9004-9.

139. Traynor BJ, Nalls M, Lai S-L, et al. Kinesin-associated protein 3 (KIFAP3) has no effect on survival in a population-based cohort of ALS patients. *Proceedings of the National Academy of Sciences* 2010; **107**(27): 12335-8.

140. Diekstra FP, van Vught PW, van Rheenen W, et al. UNC13A is a modifier of survival in amyotrophic lateral sclerosis. *Neurobiol Aging* 2012; **33**(3): 630 e3-8.

141. Kirby J, Goodall E, Smith W, et al. Broad clinical phenotypes associated with TAR-DNA binding protein (TARDBP) mutations in amyotrophic lateral sclerosis. *neurogenetics* 2010; **11**(2): 217-25.

142. Haverkamp L, Appel V, Appel S. Natural history of amyotrophic lateral sclerosis in a database population. Validation of a scoring system and a model for survival prediction. *Brain* 1995; **118**(Pt 3): 707 - 19.

143. McCombe PA, Henderson RD. Effects of gender in amyotrophic lateral sclerosis. *Gender Medicine* 2010; **7**(6): 557-70.

144. Sabatelli M, Zollino M, Luigetti M, et al. Uncovering amyotrophic lateral sclerosis phenotypes: Clinical features and long-term follow-up of upper motor neuron-dominant ALS. *Amyotrophic Lateral Sclerosis* 2011; **12**(4): 278-82.

145. Kim HY, Ki CS, Koh SH, Park KH, Sunwoo IN, Kim SH. Clinical characteristics of familial amyotrophic lateral sclerosis with a Phe20Cys mutation in the SOD1 gene in a Korean family. *Amyotrophic Lateral Sclerosis* 2007; **8**(2): 73-8.

146. Veldink JH, Bär PR, Joosten EAJ, Otten M, Wokke JHJ, van den Berg LH. Sexual differences in onset of disease and response to exercise in a transgenic model of ALS. *Neuromuscular disorders : NMD* 2003; **13**(9): 737-43.

147. Dunckley T, Huentelman MJ, Craig DW, et al. Whole-Genome Analysis of Sporadic Amyotrophic Lateral Sclerosis. *New England Journal of Medicine* 2007; **357**(8): 775-88.

148. Eisen A, Kim S, Pant B. Amyotrophic lateral sclerosis (ALS): A phylogenetic disease of the corticomotoneuron? *Muscle & Nerve* 1992; **15**(2): 219-24.

149. Chou SM, Norris FH. Issues & Opinions: Amyotrophic lateral sclerosis: Lower motor neuron disease spreading to upper motor neurons. *Muscle & Nerve* 1993; **16**(8): 864-9.

150. Ravits J, Paul P, Jorg C. Focality of upper and lower motor neuron degeneration at the clinical onset of ALS. Hagerstown, MD, ETATS-UNIS: Lippincott Williams & Wilkins; 2007.

151. Ravits J, Laurie P, Fan Y, Moore DH. Implications of ALS focality: Rostral–caudal distribution of lower motor neuron loss postmortem. *Neurology* 2007; **68**(19): 1576-82.

152. Turner MR, Modo M. Advances in the application of MRI to amyotrophic lateral sclerosis. *Expert Opinion on Medical Diagnostics* 2010; **4**(6): 483-96.

153. Verstraete E, van den Heuvel MP, Veldink JH, et al. Motor Network Degeneration in Amyotrophic Lateral Sclerosis: A Structural and Functional Connectivity Study. *PLoS ONE* 2010; **5**(10): e13664.

154. Verstraete E, Veldink JH, van den Berg LH, van den Heuvel MP. Structural brain network imaging shows expanding disconnection of the motor system in amyotrophic lateral sclerosis. *Human Brain Mapping* 2013: n/a-n/a.

155. Gargiulo-Monachelli GM, Janota F, Bettini M, Shoesmith CL, Strong MJ, Sica REP. Regional spread pattern predicts survival in patients with sporadic amyotrophic lateral sclerosis. *European Journal of Neurology* 2012; **19**(6): 834-41.

156. Ravits JM, La Spada AR. ALS motor phenotype heterogeneity, focality, and spread: deconstructing motor neuron degeneration. *Neurology* 2009; **73**(10): 805-11.

157. Dangond F, Hwang D, Camelo S, et al. Molecular signature of late-stage human ALS revealed by expression profiling of postmortem spinal cord gray matter. *Physiological Genomics* 2004; **16**: 229-39.

158. Ross CA, Poirier MA. What is the role of protein aggregation in neurodegeneration? *Nature Reviews Molecular Cell Biology* 2005; **6**(11): 891-8.

159. Polymenidou M, Cleveland Don W. The Seeds of Neurodegeneration: Prion-like Spreading in ALS. *Cell* 2011; **147**(3): 498-508.

160. Natale G, Pompili E, Biagioni F, Paparelli S, Lenzi P, Fornai F. Histochemical approaches to assess cell-to-cell transmission of misfolded proteins in neurodegenerative diseases; 2013.

161. Brooks B. The role of axonal transport in neurodegenerative disease spread: a meta-analysis of experimental and clinical poliomyelitis compares with amyotrophic lateral sclerosis. *Can J Neurol Sci* 1991; **18**(3): 435-8.

162. Munsat TLM, Andres PLMR, Finison LP, Conlon TM, Thibodeau LP. The natural history of motoneuron loss in amyotrophic lateral sclerosis. *Neurology* 1988; **38**(3): 409-13.

163. Roche JC, Rojas-Garcia R, Scott KM, et al. A proposed staging system for amyotrophic lateral sclerosis. *Brain* 2012; **135**(3): 847-52.

164. Gordon P, Cheng B, Salachas F, et al. Progression in ALS is not linear but is curvilinear. *Journal of Neurology* 2010; **257**(10): 1713-7.

165. Turner MR, Bakker M, Sham P, Shaw CE, Leigh PN, Al-Chalabi A. Prognostic modelling of therapeutic interventions in amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis* 2002; **3**(1): 15-21.

166. Chio A, Mora G, Leone M, et al. Early symptom progression rate is related to ALS outcome: A prospective population-based study. *Neurology* 2002; **59**: 99 - 103.

167. Fujimura-Kiyono C, Kimura F, Ishida S, et al. Onset and spreading patterns of lower motor neuron involvements predict survival in sporadic amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* 2011; **82**(11): 1244-9.

168. Armon C, Graves MC, Moses D, et al. Linear estimates of disease progression predict survival in patients with amyotrophic lateral sclerosis. *Muscle & Nerve* 2000; **23**(6): 874-82.

169. Saris CGJ, Groen EJN, Koekkoek JAF, Veldink JH, Van Den Berg LH. Meta-analysis of gene expression profiling in amyotrophic lateral sclerosis: A comparison between transgenic mouse models and human patients. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 2013; **14**(3): 177-89.

170. Olsen MK, Roberds SL, Ellerbrock BR, Fleck TJ, McKinley DK, Gurney ME. Disease mechanisms revealed by transcription profiling in SOD1-G93A transgenic mouse spinal cord. *Annals of Neurology* 2001; **50**(6): 730-40.

171. Perrin FE, Boisset G, Docquier M, Schaad O, Descombes P, Kato AC. No widespread induction of cell death genes occurs in pure motoneurons in an amyotrophic lateral sclerosis mouse model. *Human Molecular Genetics* 2005; **14**(21): 3309-20.

172. Yoshihara T, Ishigaki S, Yamamoto M, et al. Differential expression of inflammation- and apoptosis-related genes in spinal cords of a mutant SOD1 transgenic mouse model of familial amyotrophic lateral sclerosis. *Journal of Neurochemistry* 2002; **80**(1): 158-67.

173. Perrin FE, Boisset G, Lathuilière A, Kato AC. Cell death pathways differ in several mouse models with motoneurone disease: analysis of pure motoneurone populations at a presymptomatic age. *Journal of Neurochemistry* 2006; **98**(6): 1959-72.

174. Ferraiuolo L, Heath PR, Holden H, Kasher P, Kirby J, Shaw PJ. Microarray Analysis of the Cellular Pathways Involved in the Adaptation to and Progression of Motor Neuron Injury in the SOD1 G93A Mouse Model of Familial ALS. *The Journal of Neuroscience* 2007; **27**(34): 9201-19.

175. D'Arrigo A, Colavito D, Peña-Altamira E, et al. Transcriptional Profiling in the Lumbar Spinal Cord of a Mouse Model of Amyotrophic Lateral Sclerosis: A Role for Wild-Type Superoxide Dismutase 1 in Sporadic Disease? *Journal of Molecular Neuroscience* 2010; **41**(3): 404-15.

176. Chen H, Guo Y, Hu M, Duan W, Chang G, Li C. Differential expression and alternative splicing of genes in lumbar spinal cord of an amyotrophic lateral sclerosis mouse model. *Brain Research* 2010; **1340**(0): 52-69.

177. Fukada Y, Yasui K, Kitayama M, et al. Gene expression analysis of the murine model of amyotrophic lateral sclerosis: Studies of the Leu126delTT mutation in SOD1. *Brain Research* 2007; **1160**(0): 1-10.

178. Kudo L, Parfenova L, Vi N, et al. Integrative gene-tissue microarray-based approach for identification of human disease biomarkers: application to amyotrophic lateral sclerosis. *Hum Mol Genet* 2010; **19**(16): 3233 - 53.

179. Dangond F, Hwang D, Camelo S, et al. Molecular signature of late-stage human ALS revealed by expression profiling of postmortem spinal cord gray matter. *Physiol Genomics* 2004; **16**: 229 - 39.

180. Offen D, Barhum Y, Melamed E, Embacher N, Schindler C, Ransmayr G. Spinal Cord mRNA Profile in Patients with ALS: Comparison with Transgenic Mice Expressing the Human SOD-1 Mutant. *Journal of Molecular Neuroscience* 2009; **38**(2): 85-93.

181. Malaspina A, Kaushik N, de Belleroche J. Differential expression of 14 genes in amyotrophic lateral sclerosis spinal cord detected using gridded cDNA arrays. *J Neurochem* 2001; **77**: 132 - 45.

182. Lederer C, Torrisi A, Pantelidou M, Santama N, Cavallaro S. Pathways and genes differentially expressed in the motor cortex of patients with sporadic amyotrophic lateral sclerosis. *BMC Genomics* 2007; **8**(1): 26.

183. Al-Chalabi A, Jones A, Troakes C, King A, Al-Sarraj S, den Berg L. The genetics and neuropathology of amyotrophic lateral sclerosis. *Acta Neuropathologica* 2012; **124**(3): 339-52.

184. Wang XS, Simmons Z, Liu W, Boyer PJ, Connor JR. Differential expression of genes in amyotrophic lateral sclerosis revealed by profiling the post mortem cortex. *Amyotrophic Lateral Sclerosis* 2006; **7**(4): 201-16.

185. Kirby J, Ning K, Ferraiuolo L, et al. Phosphatase and tensin homologue/protein kinase B pathway linked to motor neuron survival in human superoxide dismutase 1-related amyotrophic lateral sclerosis. *Brain* 2011; **134**(2): 506-17.

186. Cox LE, Ferraiuolo L, Goodall EF, et al. Mutations in CHMP2B in Lower Motor Neuron Predominant Amyotrophic Lateral Sclerosis (ALS). *PLoS ONE* 2010; **5**(3): e9872.

187. Goldman JE, Yen S-H. Cytoskeletal protein abnormalities in neurodegenerative diseases. *Annals of Neurology* 1986; **19**(3): 209-23.

188. Sotelo-Silveira JR, Lepanto P, Elizondo V, et al. Axonal mitochondrial clusters containing mutant SOD1 in transgenic models of ALS. *Antioxidants & Redox Signaling* 2009; **11**(7): 1535-45.

189. Papiani G, Ruggiano A, Fossati M, et al. Restructured endoplasmic reticulum generated by mutant amyotrophic lateral sclerosis-linked VAPB is cleared by the proteasome. *Journal of Cell Science* 2012; **125**(15): 3601-11.

190. Fecto F, Siddique T. UBQLN2/P62 cellular recycling pathways in amyotrophic lateral sclerosis and frontotemporal dementia. *Muscle & Nerve* 2012; **45**(2): 157-62.

191. Eisen A, Krieger C. Pathogenic mechanisms in sporadic amyotrophic lateral sclerosis. *The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques* 1993; **20**(4): 286.

192. Jiang Y, Yamamoto M, Kobayashi Y, et al. Gene expression profile of spinal motor neurons in sporadic amyotrophic lateral sclerosis. *Ann Neurol* 2005; **57**: 236 - 51.

193. Rabin SJ, Kim JMH, Baughn M, et al. Sporadic ALS has compartment-specific aberrant exon splicing and altered cell–matrix adhesion biology. *Human Molecular Genetics* 2010; **19**(2): 313-28.

194. Mougeot J-L, Li Z, Price A, Wright F, Brooks B. Microarray analysis of peripheral blood lymphocytes from ALS patients and the SAFE detection of the KEGG ALS pathway. *BMC Medical Genomics* 2011; **4**(1): 74.

195. Zhang R, Hadlock K, Do H, et al. Gene expression profiling in peripheral blood mononuclear cells from patients with sporadic amyotrophic lateral sclerosis (sALS). *J Neuroimmunol* 2010; **230**(1-2): 114 - 23.

196. Pradat PF, Dubourg O, de Tapia M, et al. Muscle Gene Expression Is a Marker of Amyotrophic Lateral Sclerosis Severity. *Neurodegenerative Diseases* 2012; **9**(1): 38-52.

197. Shtilbans A, Choi S-G, Fowkes ME, et al. Differential gene expression in patients with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis* 2011; **0**(0): 1-7.

198. Saris C, Horvath S, van Vught P, et al. Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics* 2009; **10**: 405.

199. Cooper-Knock J, Kirby J, Ferraiuolo L, Heath P, Rattray M, Shaw P. Gene expression profiling in human neurodegenerative disease. *Nature reviews Neurology* 2012; **8**(9): 518-30.

200. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet* 2010; **6**(4): e1000888.

201. Diekstra FP, Saris CGJ, van Rheenen W, et al. Mapping of Gene Expression Reveals <italic>CYP27A1</italic> as a Susceptibility Gene for Sporadic ALS. *PLoS ONE* 2012; **7**(4): e35333.

202. Gallus GN, Dotti MT, Federico A. Clinical and molecular diagnosis of cerebrotendinous xanthomatosis with a review of the mutations in the CYP27A1 gene. *Neurol Sci* 2006; **27**(2): 143-9.

203. Cali JJ, Hsieh CL, Francke U, Russell DW. Mutations in the bile acid biosynthetic enzyme sterol 27-hydroxylase underlie cerebrotendinous xanthomatosis. *Journal of Biological Chemistry* 1991; **266**(12): 7779-83.

204. Gellera C, Ticozzi N, Pensato V, et al. ATAXIN2 CAG-repeat length in Italian patients with amyotrophic lateral sclerosis: risk factor or variant phenotype? Implication for genetic testing and counseling. *Neurobiology of Aging* 2012; **33**(8): 1847.e15-.e21.

205. Van Damme P, Veldink JH, van Blitterswijk M, et al. Expanded ATXN2 CAG repeat size in ALS identifies genetic overlap between ALS and SCA2. *Neurology* 2011; **76**(24): 2066-72.

206. Yu Z, Zhu Y, Chen-Plotkin AS, et al. PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. *PLoS ONE* 2011; **6**(3): e17951.

207. Ross OA, Rutherford NJ, Baker M, et al. Ataxin-2 repeat-length variation and neurodegeneration. *Hum Mol Genet* 2011; **20**(16): 3207-12.

208. Daoud H, Belzil V, Martins S, et al. Association of long ATXN2 CAG repeat sizes with increased risk of amyotrophic lateral sclerosis. *Arch Neurol* 2011; **68**(6): 739-42.

209. Chen Y, Huang R, Yang Y, et al. Ataxin-2 intermediate-length polyglutamine: a possible risk factor for Chinese patients with amyotrophic lateral sclerosis. *Neurobiol Aging* 2011; **32**(10): 1925 e1-5.

210. Gispert S, Kurz A, Waibel S, et al. The modulation of Amyotrophic Lateral Sclerosis risk by ataxin-2 intermediate polyglutamine expansions is a specific effect. *Neurobiol Dis* 2012; **45**(1): 356-61.

211. Van Langenhove T, van der Zee J, Engelborghs S, et al. Ataxin-2 polyQ expansions in FTLD-ALS spectrum disorders in Flanders-Belgian cohorts. *Neurobiol Aging* 2011.

212. Liu X, Lu M, Tang L, Zhang N, Chui D, Fan D. ATXN2 CAG repeat expansions increase the risk for Chinese patients with amyotrophic lateral sclerosis. *Neurobiology of Aging* 2013.

213. Lahut S, Ömür Ö, Uyan Ö, et al. ATXN2 and Its Neighbouring Gene SH2B3 Are Associated with Increased ALS Risk in the Turkish Population. *PLoS ONE* 2012; **7**(8): e42956.

214. Simpson CL, Lemmens R, Miskiewicz K, et al. Variants of the elongator protein 3 (ELP3) gene are associated with motor neuron degeneration. *Hum Mol Genet* 2009; **18**(3): 472-81.

215. Staats KA, Van Helleputte L, Jones AR, et al. Genetic ablation of phospholipase C delta 1 increases survival in SOD1G93A mice. *Neurobiology of Disease* 2013; (0).

216. Lewis CM, Ng MY, Butler AW, et al. Genome-Wide Association Study of Major Recurrent Depression in the U.K. Population. *American Journal of Psychiatry* 2010; **167**(8): 949-57.

217. Gaysina D, Cohen-Woods S, Chow PC, et al. Association of the dystrobrevin binding protein 1 gene (DTNBP1) in a bipolar case–control study (BACCS). *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2009; **150B**(6): 836-44.

218. Landers JE, Melki J, Meininger V, et al. Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proc Natl Acad Sci U S A* 2009; **106**(22): 9004-9.

219. van Es M, Veldink J, Saris C, et al. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat Genet* 2009; **41**(10): 1083 - 7.

220. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American journal of human genetics* 2007; **81**(3): 559-75.

221. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**(8): 904-9.

222. Van Langenhove T, van der Zee J, Engelborghs S, et al. Ataxin-2 polyQ expansions in FTLD-ALS spectrum disorders in Flanders-Belgian cohorts. *Neurobiology of Aging* 2012; **33**(5): 1004.e17-.e20.

223. Lee T, Li YR, Ingre C, et al. Ataxin-2 intermediate-length polyglutamine expansions in European ALS patients. *Human Molecular Genetics* 2011; **20**(9): 1697-700.

224. Sorarù G, Clementi M, Forzan M, et al. ALS risk but not phenotype is affected by ataxin-2 intermediate length polyglutamine expansion. *Neurology* 2011; **76**(23): 2030-1.

225. Smith BN, Newhouse S, Shatunov A, et al. The C9ORF72 expansion mutation is a common cause of ALS+/-FTD in Europe and has a single founder. *Eur J Hum Genet* 2012.

226. Jones AR, Woollacott I, Shatunov A, et al. Residual association at C9orf72 suggests an alternative amyotrophic lateral sclerosis-causing hexanucleotide repeat. *Neurobiology of aging* 2013; **34**(9): 2234.e1-.e7.

227. Beck J, Poulter M, Hensman D, et al. Large C9orf72 Hexanucleotide Repeat Expansions Are Seen in Multiple Neurodegenerative Syndromes and Are More Frequent Than Expected in the UK Population. *American journal of human genetics* 2013.

228. Liang L, Morar N, Dixon AL, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Research* 2013; **23**(4): 716-26.

229. Johnston C, Stanton B, Turner M, et al. Amyotrophic lateral sclerosis in an urban setting. *Journal of Neurology* 2006; **253**(12): 1642-3.

230. Chio A, Mora G, Calvo A, Mazzini L, Bottacchi E, Mutani R. Epidemiology of ALS in Italy: a 10-year prospective population-based study. *Neurology* 2009; **72**(8): 725-31.

231. Rosen D, Siddique T, Patterson D, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 1993; **362**: 59 - 62.

232. Kabashi E, Valdmanis PN, Dion P, et al. TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat Genet* 2008; **40**(5): 572-4.

233. Greenway MJ, Andersen PM, Russ C, et al. ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat Genet* 2006; **38**(4): 411-3.

234. Chow CY, Landers JE, Bergren SK, et al. Deleterious variants of FIG4, a phosphoinositide phosphatase, in patients with ALS. *American journal of human genetics* 2009; **84**(1): 85-8.

235. Wu C-H, Fallini C, Ticozzi N, et al. Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature* 2012; **488**(7412): 499-503.

236. Byrne S, Hardiman O. Familial aggregation in amyotrophic lateral sclerosis. *Annals of Neurology* 2010; **67**(4): 554-.

237. Hanby MF, Scott KM, Scotton W, et al. The risk to relatives of patients with sporadic amyotrophic lateral sclerosis. *Brain* 2011; **134**(Pt 12): 3454-7.

238. Al-Chalabi A, Lewis CM. Modelling the effects of penetrance and family size on rates of sporadic and familial disease. *Human heredity* 2011; **71**(4): 281-8.

239. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 2008; **18**(11): 1851-8.

240. Corrado L, Mazzini L, Oggioni G, et al. ATXN-2 CAG repeat expansions are interrupted in ALS patients. *Human Genetics* 2011; **130**(4): 575-80.

241. Renčiuk D, Zemánek M, Kejnovská I, Vorlíčková M. Quadruplex-forming properties of FRAXA (CGG) repeats interrupted by (AGG) triplets. *Biochimie* 2009; **91**(3): 416-22.

242. Pickering-Brown S, Rollinson PS, Snowden J, Gerhard A, Neary D, Mann D. FTLD Repeat Expansions In C9orf72:Evidence For Variability In The Repeat Sequence *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 2012; **8**(4): 93.

243. Inoue K, Zhuang L, Ganapathy V. Human Na+-coupled citrate transporter: primary structure, genomic organization, and transport function. *Biochemical and Biophysical Research Communications* 2002; **299**(3): 465-71.

244. Zubenko GS, Hughes HB, Zubenko WN. D10S1423 identifies a susceptibility locus for Alzheimer's disease (AD7) in a prospective, longitudinal, double-blind study of asymptomatic individuals: Results at 14 years. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2010; **153B**(2): 359-64.

245. Schymick JC, Scholz SW, Fung H-C, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *The Lancet Neurology* 2007; **6**(4): 322-8.

246. Rutherford NJB, DeJesus- Hernandez MBS, Baker MCB, et al. C9ORF72 hexanucleotide repeat expansions in patients with ALS from the Coriell Cell Repository. *Neurology* 2012; **79**(5): 482-3.

247. Gomez-Tortosa EMDP, Gallego JP, Guerrero-Lopez RP, et al. C9ORF72 hexanucleotide expansions of 20-22 repeats are associated with frontotemporal deterioration. *Neurology* 2013; **80**(4): 366-70.

248. Kwee LC, Liu Y, Haynes C, et al. A high-density genome-wide association screen of sporadic ALS in US veterans. *PLoS ONE* 2012; **7**(3): e32768.

249. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; **25**(9): 1105-11.

250. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 2009; **10**(3): R25.

251. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**(16): 2078-9.

252. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 2010; **20**(9): 1297-303.

253. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics* 2011; **27**(15): 2156-8.

254. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 2011.

255. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research* 2010.

256. Mok K, Traynor BJ, Schymick J, et al. The chromosome 9 ALS and FTD locus is probably derived from a single founder. *Neurobiology of Aging* 2012; **33**(1): 209.e3-.e8.

257. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 2009; **4**(1): 44-57.

258. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**(1): 1-13.

259. Jensen LJ, Kuhn M, Stark M, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 2009; **37**(Database issue): D412-6.

260. Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research* 2010; **38**(suppl 2): W214-W20.

261. Liekens AM, De Knijf J, Daelemans W, Goethals B, De Rijk P, Del-Favero J. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol* 2011; **12**(6): R57.

262. Aoki M, Rothstein JD, Lin CLG, et al. Mutations of the glutamate transporter gene EAAT2 do not cause the abnormal EAAT2 transcripts in patients with amyotrophic lateral sclerosis. *Neuromuscular Disorders* 1997; **7**(6): 475-.

263. Edri-Brami M, Rosental B, Hayoun D, et al. Glycans in Sera of Amyotrophic Lateral Sclerosis Patients and Their Role in Killing Neuronal Cells. *PLoS ONE* 2012; **7**(5): e35772.

264. Tanaka H, Shimazawa M, Kimura M, et al. The potential of GPNMB as novel neuroprotective factor in amyotrophic lateral sclerosis. *Scientific reports* 2012; **2**: 573.

265. Lüdemann N, Clement A, Hans VH, Leschik J, Behl C, Brandt R. O-Glycosylation of the Tail Domain of Neurofilament Protein M in Human Neurons and in Spinal Cord Tissue of a Rat

Model of Amyotrophic Lateral Sclerosis (ALS). *Journal of Biological Chemistry* 2005; **280**(36): 31648-58.

266. Shan X, Vocadlo DJ, Krieger C. Reduced protein O-glycosylation in the nervous system of the mutant SOD1 transgenic mouse model of amyotrophic lateral sclerosis. *Neuroscience Letters* 2012; **516**(2): 296-301.

267. Nishimura AL, Mitne-Neto M, Silva HC, et al. A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *Am J Hum Genet* 2004; **75**(5): 822-31.

268. Al-Saif A, Al-Mohanna F, Bohlega S. A mutation in sigma-1 receptor causes juvenile amyotrophic lateral sclerosis. *Annals of Neurology* 2011; **70**(6): 913-9.

269. Traynor BJ, Nalls M, Lai SL, et al. Kinesin-associated protein 3 (KIFAP3) has no effect on survival in a population-based cohort of ALS patients. *Proc Natl Acad Sci U S A* 2010; **107**(27): 12335-8.

270. Pantelidou M, Zographos SE, Lederer CW, Kyriakides T, Pfaffl MW, Santama N. Differential expression of molecular motors in the motor cortex of sporadic ALS. *Neurobiology of Disease* 2007; **26**(3): 577-89.

271. Alexianu ME, Ho B-K, Mohamed AH, La Bella V, Smith RG, Appel SH. The role of calcium-binding proteins in selective motoneuron vulnerability in amyotrophic lateral sclerosis. *Annals of Neurology* 1994; **36**(6): 846-58.

272. Siklos L, Engelhardt J, Harati Y, Smith R, Joo F, Appel S. Ultrastructural evidence for altered calcium in motor nerve terminals in amyotropic lateral sclerosis. *Ann Neurol* 1996; **39**: 203 - 16.

273. Siklós L, Engelhardt JI, Alexianu ME, Gurney ME, Siddique T, Appel SH. Intracellular Calcium Parallels Motoneuron Degeneration in SOD-1 Mutant Mice. *Journal of Neuropathology & Experimental Neurology* 1998; **57**(6): 571-87.

274. Sucher NJ, Lei SZ, Lipton SA. Calcium channel antagonists attenuate NMDA receptor-mediated neurotoxicity of retinal ganglion cells in culture. *Brain Research* 1991; **551**(1–2): 297-302.

275. Dykens JA. Isolated Cerebral and Cerebellar Mitochondria Produce Free Radicals when Exposed to Elevated Ca2+ and Na+: Implications for Neurodegeneration. *Journal of Neurochemistry* 1994; **63**(2): 584-91.

276. Kim SH, Zhan L, Hanson KA, Tibbetts RS. High-content RNAi screening identifies the Type 1 inositol triphosphate receptor as a modifier of TDP-43 localization and neurotoxicity. *Human Molecular Genetics* 2012; **21**(22): 4845-56.

277. Gotoh F, Kitamura A, Koto A, Kataoka K, Atsuji H. Abnormal insulin secretion in amyotrophic lateral sclerosis. *Journal of the neurological sciences* 1972; **16**(2): 201-7.

278. HIBBAR MD, DAVIDSON MB, ROSENBERG CS. Insulin Antagonism Is Not a Primary Abnormality of Amyotrophic Lateral Sclerois but Is Related to Disease Severity. *Journal of Clinical Endocrinology & Metabolism* 1986; **63**(1): 41-6.

279. Przedborski S, Donaldson DM, Murphy PL, et al. Blood Superoxide Dismutase, Catalase and Glutathione Peroxidase Activities in Familial and Sporadic Amyotrophic Lateral Sclerosis. *Neurodegeneration* 1996; **5**(1): 57-64.

280. Wilczak N, de Vos RAI, De Keyser J. Free insulin-like growth factor (IGF)-I and IGF binding proteins 2, 5, and 6 in spinal motor neurons in amyotrophic lateral sclerosis. *The Lancet* 2003; **361**(9362): 1007-11.

281. Chung YH, Joo KM, Shin CM, et al. Immunohistochemical study on the distribution of insulin-like growth factor I (IGF-I) receptor in the central nervous system of SOD1G93A mutant transgenic mice. *Brain Research* 2003; **994**(2): 253-9.

282. Zhang T, Mullane PC, Periz G, Wang J. TDP-43 neurotoxicity and protein aggregation modulated by heat shock factor and insulin/IGF-1 signaling. *Human Molecular Genetics* 2011; **20**(10): 1952-65.

283. Mitsuma T, Nogimori T, Adachi K, Mukoyama M, Ando K. Concentrations of Immunoreactive Thyrotropin-Releasing Hormone in Spinal Cord of Patients with Amyotrophic Lateral Sclerosis. *American Journal of the Medical Sciences March/April* 1984; **287**(2): 34-6.

284. Imoto K, Saida K, Iwamura K, Saida T, Nishitani H. Amyotrophic lateral sclerosis: a double-blind crossover trial of thyrotropin-releasing hormone. *Journal of neurology, neurosurgery, and psychiatry* 1984; **47**(12): 1332-4.

285. Brooke MH, Florence JM, Heller SL, et al. Controlled trial of thyrotropin releasing hormone in amyotrophic lateral sclerosis. *Neurology* 1986; **36**(2): 146.

286. Caroscio JT, Cohen JA, Zawodniak J, et al. A double-blind, placebo-controlled trial of TRH in amyotrophic lateral sclerosis. *Neurology* 1986; **36**(2): 141.

287. Hentati A, Bejaoui K, Pericak-Vance M, et al. Linkage of recessive familial amyotrophic lateral sclerosis to chromosome 2q33-q35. *Nat Genet* 1994; **7**: 425 - 8.

288. Devon RS, Helm JR, Rouleau GA, et al. The first nonsense mutation in alsin results in a homogeneous phenotype of infantile-onset ascending spastic paralysis with bulbar involvement in two siblings. *Clinical Genetics* 2003; **64**(3): 210-5.

289. Johnson J, Mandrioli J, Benatar M, et al. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* 2010; **68**(5): 857 - 64.

290. Deng H, Chen W, Hong S, et al. Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature* 2011; **477**(7363): 211 - 5.

291. Ticozzi N, LeClerc AL, Keagle PJ, et al. Paraoxonase gene mutations in amyotrophic lateral sclerosis. *Annals of Neurology* 2010; **68**(1): 102-7.

292. Morahan JM, Yu B, Trent RJ, Pamphlett R. A gene–environment study of the paraoxonase 1 gene and pesticides in amyotrophic lateral sclerosis. *NeuroToxicology* 2007; **28**(3): 532-40.

293. Cronin S, Greenway MJ, Prehn JHM, Hardiman O. Paraoxonase promoter and intronic variants modify risk of sporadic amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* 2007; **78**(9): 984-6.

294. Landers JE, Shi L, Cho TJ, et al. A common haplotype within the PON1 promoter region is associated with sporadic ALS. *Amyotrophic lateral sclerosis : official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* 2008; **9**(5): 306-14.

295. van Dijk JP, Schelhaas HJ, Van Schaik IN, Janssen HMHA, Stegeman DF, Zwarts MJ. Monitoring disease progression using high-density motor unit number estimation in amyotrophic lateral sclerosis. *Muscle & Nerve* 2010; **42**(2): 239-44.

296. Nutini M, Spalloni A, Florenzano F, et al. Increased expression of the beta3 subunit of voltage-gated Na+ channels in the spinal cord of the SOD1G93A mouse. *Molecular and Cellular Neuroscience* 2011; **47**(2): 108-18.

297. Puls I, Jonnakuty C, LaMonte B, et al. Mutant dynactin in motor neuron disease. *Nat Genet* 2003; **33**: 455 - 6.

298. Shoichet SA, Waibel S, Endruhn S, et al. Identification of candidate genes for sporadic amyotrophic lateral sclerosis by array comparative genomic hybridization. *Amyotroph Lateral Scler* 2009; **10**(3): 162-9.

299. Chi C, Tian R, Liu H, et al. Follow-up study of abnormal biological indicators and gene expression in the peripheral blood of three accidentally exposed persons. *Journal of Radiation Research* 2013; **54**(5): 840-51.

300. Kanai Y, Clémençon B, Simonin A, et al. The SLC1 high-affinity glutamate and neutral amino acid transporter family. *Molecular Aspects of Medicine* 2013; **34**(2–3): 108-20.

301. Arriza JL, Eliasof S, Kavanaugh MP, Amara SG. Excitatory amino acid transporter 5, a retinal glutamate transporter coupled to a chloride conductance. *Proc Natl Acad Sci U S A* 1997; **94**(8): 4155-60.

302. Bristol LA, Rothstein JD. Glutamate transporter gene expression in amyotrophic lateral sclerosis motor cortex. *Annals of Neurology* 1996; **39**(5): 676-9.

303. Lin C-LG, Bristol LA, Jin L, et al. Aberrant RNA Processing in a Neurodegenerative Disease: the Cause for Absent EAAT2, a Glutamate Transporter, in Amyotrophic Lateral Sclerosis. *Neuron* 1998; **20**(3): 589-602.

304. Diaper DC, Adachi Y, Lazarou L, et al. Drosophila TDP-43 dysfunction in glia and muscle cells cause cytological and behavioural phenotypes that characterize ALS and FTLD. *Human Molecular Genetics* 2013; **22**(19): 3883-93.

305. Guo H, Lai L, Butchbach MER, et al. Increased expression of the glial glutamate transporter EAAT2 modulates excitotoxicity and delays the onset but not the outcome of ALS in mice. *Human Molecular Genetics* 2003; **12**(19): 2519-32.

306. Gao X, Xu Z. Mechanisms of action of angiogenin. *Acta Biochimica et Biophysica Sinica* 2008; **40**(7): 619-24.

307. Neumann M, Sampathu DM, Kwong LK, et al. Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. *Science* 2006; **314**(5796): 130-3.

308. Sreedharan J, Blair IP, Tripathi VB, et al. TDP-43 Mutations in Familial and Sporadic Amyotrophic Lateral Sclerosis. *Science* 2008; **319**(5870): 1668-72.

309. Nishimura AL, Mitne-Neto M, Silva HCA, et al. A Mutation in the Vesicle-Trafficking Protein VAPB Causes Late-Onset Spinal Muscular Atrophy and Amyotrophic Lateral Sclerosis. *American journal of human genetics* 2004; **75**(5): 822-31.

310. Lewis ME, Neff NT, Contreras PC, et al. Insulin-like Growth Factor-I: Potential for Treatment of Motor Neuronal Disorders. *Experimental Neurology* 1993; **124**(1): 73-88.

311. Allen RG. Oxidative stress and superoxide dismutase in development, aging and gene regulation. *Age* 1998; **21**(2): 47-76.

312. Festoff BW. Neuromuscular junction macromolecules in the pathogenesis of amyotrophic lateral sclerosis. *Medical Hypotheses* 1980; **6**(2): 121-31.

313. Rao JS, Hantaï D, Festoff BW. Thrombospondin, a platelet α-granule and matrix glycoprotein, is increased in muscle basement membrane of patients with amyotrophic lateral sclerosis. *Journal of the Neurological Sciences* 1992; **113**(1): 99-107.

314. Niebroj-Dobosz I, Mickielewicz A, Rowiñska-Marciñska K, Kwieciñski H. Identification of Gal(β1–3)GalNAc bearing glycoproteins in cerebrospinal fluid of amyotrophic lateral sclerosis (ALS) patients. *European Journal of Neurology* 2000; **7**(6): 679-83.

315. Palma A, de Carvalho M, Barata N, et al. Biochemical characterization of plasma in amyotrophic lateral sclerosis: Amino acid and protein composition. *Amyotrophic Lateral Sclerosis* 2005; **6**(2): 104-10.

316. Milane A, Fernandez C, Dupuis L, et al. P-glycoprotein expression and function are increased in an animal model of amyotrophic lateral sclerosis. *Neuroscience Letters* 2010; **472**(3): 166-70.

317. Appel SH, Smith RG, Engelhardt JI, Stefani E. Evidence for autoimmunity in amyotrophic lateral sclerosis. *J Neurol Sci* 1993; **118**(2): 169-74.

318. Ling Q, Jacovina A, Deora A, et al. Annexin II regulates fibrin homeostasis and neoangiogenesis in vivo. *J Clin Invest* 2004; **113**: 38 - 48.

319. Kato T, Hirano A, Kurland LT. Asymmetric involvement of the spinal cord involving both large and small anterior horn cells in a case of familial amyotrophic lateral sclerosis. *Clin Neuropathol* 1987; **6**(2): 67-70.

320. Troost D, Oord J, de Jong J, Swaab D. Lymphocytic infiltration in the spinal cord of patients with amyotrophic lateral sclerosis. *Clin Neuropathol* 1989; **8**: 289 - 94.

321. Engelhardt JI, Tajti J, Appel SH. Lymphocytic infiltrates in the spinal cord in amyotrophic lateral sclerosis. *Arch Neurol* 1993; **50**(1): 30-6.

322. Mantovani S, Garbelli S, Pasini A, et al. Immune system alterations in sporadic amyotrophic lateral sclerosis patients suggest an ongoing neuroinflammatory process. *J Neuroimmunol* 2009; **210**(1-2): 73-9.

323. Graves MC, Fiala M, Dinglasan LA, et al. Inflammation in amyotrophic lateral sclerosis spinal cord and brain is mediated by activated macrophages, mast cells and T cells. *Amyotroph Lateral Scler Other Motor Neuron Disord* 2004; **5**(4): 213-9.

324. Chiu IM, Chen A, Zheng Y, et al. T lymphocytes potentiate endogenous neuroprotective inflammation in a mouse model of ALS. *Proc Natl Acad Sci U S A* 2008; **105**(46): 17913-8.

325. Beers DR, Henkel JS, Zhao W, et al. Endogenous regulatory T lymphocytes ameliorate amyotrophic lateral sclerosis in mice and correlate with disease progression in patients with amyotrophic lateral sclerosis. *Brain* 2011; **134**(Pt 5): 1293-314.

326. Rezai-Zadeh K, Gate D, Town T. CNS infiltration of peripheral immune cells: D-Day for neurodegenerative disease? *Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology* 2009; **4**(4): 462-75.

327. Tada S, Okuno T, Yasui T, et al. Deleterious effects of lymphocytes at the early stage of neurodegeneration in an animal model of amyotrophic lateral sclerosis. *Journal of neuroinflammation* 2011; **8**(1): 19.

328. Henkel JS, Beers DR, Zhao W, Appel SH. Microglia in ALS: the good, the bad, and the resting. *Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology* 2009; **4**(4): 389-98.

329. Evans MC, Couch Y, Sibson N, Turner MR. Inflammation and neurovascular changes in amyotrophic lateral sclerosis. *Molecular and cellular neurosciences* 2013; **53**: 34-41.

330. Haraldsen G, Kvale D, Lien B, Farstad IN, Brandtzaeg P. Cytokine-regulated expression of E-selectin, intercellular adhesion molecule-1 (ICAM-1), and vascular cell adhesion molecule-1 (VCAM-1) in human microvascular endothelial cells. *Journal of immunology (Baltimore, Md : 1950)* 1996; **156**(7): 2558-65.

331. Shimizu Y, Newman W, Gopal TV, et al. Four molecular pathways of T cell adhesion to endothelial cells: roles of LFA-1, VCAM-1, and ELAM-1 and changes in pathway hierarchy under different activation conditions. *J Cell Biol* 1991; **113**(5): 1203-12.

332. Verbeek MM, Westphal JR, Ruiter DJ, de Waal RM. T lymphocyte adhesion to human brain pericytes is mediated via very late antigen-4/vascular cell adhesion molecule-1 interactions. *Journal of immunology (Baltimore, Md : 1950)* 1995; **154**(11): 5876-84.

333. Veldink JH, van den Berg LH, Cobben JM, et al. Homozygous deletion of the survival motor neuron 2 gene is a prognostic factor in sporadic ALS. *Neurology* 2001; **56**(6): 749-52.

334. Veldink JH, Kalmijn S, Van der Hout AH, et al. SMN genotypes producing less SMN protein increase susceptibility to and severity of sporadic ALS. *Neurology* 2005; **65**(6): 820-5.

335. Gavrilov DK, Shi X, Das K, Gilliam TC, Wang CH. Differential SMN2 expression associated with SMA severity. *Nat Genet* 1998; **20**(3): 230-1.

336. Zou T, Ilangovan R, Yu F, Xu Z, Zhou J. SMN protects cells against mutant SOD1 toxicity by increasing chaperone activity. *Biochemical and Biophysical Research Communications* 2007; **364**(4): 850-5.

337. Turner BJ, Parkinson NJ, Davies KE, Talbot K. Survival motor neuron deficiency enhances progression in an amyotrophic lateral sclerosis mouse model. *Neurobiology of Disease* 2009; **34**(3): 511-7.

338. Yoshino Y, Koike H, Akai K. Free amino acids in motor cortex of amyotrophic lateral sclerosis. *Experientia* 1979; **35**(2): 219-20.

339. Perry TL, Hansen S, Jones K. Brain glutamate deficiency in amyotrophic lateral sclerosis. *Neurology* 1987; **37**(12): 1845-8.

340. Krieger C, Perry TL, Hansen S, Mitsumoto H. The wobbler mouse: amino acid contents in brain and spinal cord. *Brain Res* 1991; **551**(1-2): 142-4.

341. Kuroda K. [Effects of excitatory sulfur amino acids on glutamate transport in synaptosomes isolated from the rat cerebral cortex]. *Rinsho shinkeigaku = Clinical neurology* 1998; **38**(12): 1019-23.

342. Chen D, Shen L, Wang L, et al. Association of polymorphisms in vascular endothelial growth factor gene with the age of onset of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis* 2007; **8**(3): 144-9.

343. Tolosa L, Mir M, Asensio VJ, Olmos G, Lladó J. Vascular endothelial growth factor protects spinal cord motoneurons against glutamate-induced excitotoxicity via phosphatidylinositol 3-kinase. *Journal of Neurochemistry* 2008; **105**(4): 1080-90.

344. Gray A, Stewart I, Tenesa A. Advanced Complex Trait Analysis. *Bioinformatics* 2012; **28**(23): 3134-6.

345. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. *American journal of human genetics* 2011; **88**(1): 76-82.

346. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007; **23**(10): 1294-6.

347. Carrì MT, Ferri A, Casciati A, Celsi F, Ciriolo MR, Rotilio G. Copper-dependent oxidative stress, alteration of signal transduction and neurodegeneration in amyotrophic lateral sclerosis. *Funct Neurol* 2001; **16**(4 Suppl): 181-8.

348. Narai H, Nagano I, Ilieva H, et al. Prevention of spinal motor neuron death by insulin-like growth factor-1 associating with the signal transduction systems in SODG93A transgenic mice. *Journal of Neuroscience Research* 2005; **82**(4): 452-7.

349. Seilhean D, Cazeneuve C, Thuriès V, et al. Accumulation of TDP-43 and α-actin in an amyotrophic lateral sclerosis patient with the K17I ANG mutation. *Acta Neuropathologica* 2009; **118**(4): 561-73.

350. Murray M, DeJesus-Hernandez M, Rutherford N, et al. Clinical and neuropathologic heterogeneity of c9FTD/ALS associated with hexanucleotide repeat expansion in C9ORF72. *Acta Neuropathologica* 2011; **122**(6): 673-90.

351. Jones DTW, Kocialkowski S, Liu L, et al. Tandem Duplication Producing a Novel Oncogenic BRAF Fusion Gene Defines the Majority of Pilocytic Astrocytomas. *Cancer Research* 2008; **68**(21): 8673-7.

352. Al-Chalabi A, Kwak S, Mehler M, et al. Genetic and epigenetic studies of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 2013; **14**(S1): 44-52.

353. Talbot K. Motor neuron disease: THE BARE ESSENTIALS. *Practical Neurology* 2009; **9**(5): 303-9.

354. Veldink JH, Wokke JHJ, van der Wal G, Vianney de Jong JMB, van den Berg LH. Euthanasia and Physician-Assisted Suicide among Patients with Amyotrophic Lateral Sclerosis in the Netherlands. *New England Journal of Medicine* 2002; **346**(21): 1638-44.

355. Johnston CA, Stanton BR, Turner MR, et al. Amyotrophic lateral sclerosis in an urban setting: a population based study of inner city London. *J Neurol* 2006; **253**(12): 1642-3.

356. Steen IVD, Berg J-PVD, Buskens E, Lindeman E, Van Den Berg LH. The costs of amyotrophic lateral sclerosis, according to type of care. *Amyotrophic Lateral Sclerosis* 2009; **10**(1): 27-34.

357. Klein LM, Forshew DABSNRN. The economic impact of ALS. *Neurology* 1996; **47(4) Supplement**(2): 126S-9S.

358. Zinman L, Cudkowicz M. Emerging targets and treatments in amyotrophic lateral sclerosis. *The Lancet Neurology* 2011; **10**(5): 481-90.

359. NICE. Guide to the methods of technology appraisal2008. (accessed May 2013).

360. EuroQol. EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy* 1990; **16**(3): 199-208.

361. Winter Y, Schepelmann K, Spottke A, et al. Health-related quality of life in ALS, myasthenia gravis and facioscapulohumeral muscular dystrophy. *Journal of Neurology* 2010; **257**(9): 1473-81.

362. UKMND-LiCALS GS. Lithium in patients with amyotrophic lateral sclerosis (LiCALS): a phase 3 multicentre, randomised, double-blind, placebo-controlled trial. *The Lancet Neurology* 2013; **12**(4): 339-45.

363. Al-Chalabi A, Shaw P, Young C, et al. Protocol for a double-blind randomised placebo-controlled trial of lithium carbonate in patients with amyotrophic Lateral Sclerosis (LiCALS) [Eudract number: 2008-006891-31]. *BMC Neurology* 2011; **11**(1): 111.

364. Dolan P, Stalmeier P. The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. *Journal of Health Economics* 2003; **22**(3): 445-58.

365. Greiner W, Claes C, Busschbach JJV, Schulenburg JM. Validating the EQ-5D with time trade off for the German population. *Eur J Health Econ* 2005; **6**(2): 124-30.

366. Dolan P. Modelling valuations for health states: the effect of duration. *Health policy (Amsterdam, Netherlands)* 1996; **38**(3): 189-203.

367. Gibbons C, Mills R, Thornton E, et al. Rasch analysis of the Hospital Anxiety and Depression Scale (HADS) for use in motor neurone disease. *Health and Quality of Life Outcomes* 2011; **9**(1): 82.

368. Kiebert GM, Green C, Murphy C, et al. Patients' health-related quality of life and utilities associated with different stages of amyotrophic lateral sclerosis. *Journal of the neurological sciences* 2001; **191**(1): 87-93.

369. Norquist JM, Jenkinson C, Fitzpatrick R, Swash M. Factors which predict physical and mental health status in patients with amyotrophic lateral sclerosis over time. *Amyotrophic Lateral Sclerosis* 2003; **4**(2): 112-7.

370. Chiò A, Gauthier A, Montuschi A, et al. A cross sectional study on determinants of quality of life in ALS. *Journal of Neurology, Neurosurgery & Psychiatry* 2004; **75**(11): 1597-601.

371. Jenkinson C, Fitzpatrick R, Swash M, Peto V. The ALS Health Profile Study: quality of life of amyotrophic lateral sclerosis patients and carers in Europe. *Journal of Neurology* 2000; **247**(11): 835-40.

372. De groot IJM, Post MWM, Heuveln Tv, Van den berg LH, Lindeman E. Cross-sectional and longitudinal correlations between disease progression and different health-related quality of life domains in persons with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis* 2007; **8**(6): 356-61.

373. Averill A, Kasarskis E, Segerstrom S. Psychological health in patients with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis* 2007; **8**: 243 - 54.

374. Dolan P. Using Happiness to Value Health. London, UK: Office of Health Economics. 2011; (Report No. 1).

375. Jensen MP, Abresch RT, Carter GT, McDonald CM. Chronic Pain in Persons With Neuromuscular Disease. *Archives of physical medicine and rehabilitation* 2005; **86**(6): 1155-63.

376. Behari M, Srivastava AK, Pandey RM. Quality of life in patients with Parkinson's disease. *Parkinsonism & related disorders* 2005; **11**(4): 221-6.

377. Ho AK, Gilbert AS, Mason SL, Goodman AO, Barker RA. Health-related quality of life in Huntington's disease: Which factors matter most? *Movement Disorders* 2009; **24**(4): 574-8.

378. Oh H, Sin M-K, Schepp KG, Choi-Kwon S. Depressive Symptoms and Functional Impairment Among Amyotrophic Lateral Sclerosis Patients in South Korea. *Rehabilitation Nursing* 2012; **37**(3): 136-44.

379. Rabkin JG, Wagner GJ, Del Bene M. Resilience and Distress Among Amyotrophic Lateral Sclerosis Patients and Caregivers. *Psychosomatic Medicine* 2000; **62**(2): 271-9.

380. Lillo P, Mioshi E, Zoing MC, Kiernan MC, Hodges JR. How common are behavioural changes in amyotrophic lateral sclerosis? *Amyotrophic Lateral Sclerosis* 2011; **12**(1): 45-51.

381. Mendez MF, Perryman KM. Neuropsychiatric Features of Frontotemporal DementiaEvaluation of Consensus Criteria and Review. *The Journal of Neuropsychiatry and Clinical Neurosciences* 2002; **14**(4): 424-9.

Appendices

Tables Appendix

Chapter 2

Name (Gene/SNP)	Associated Allele	Case, Control Ratios	Chi Square	p-value
SH2B3/rs2239194	G	0.921, 0.867	3.457	0.063
SH2B3/rs3184504	A	0.396, 0.394	0.001	0.98
SH2B3/rs739496	G	0.329, 0.298	0.49	0.4838
ATXN2/rs10849949	G	0.330, 0.298	0.54	0.4623
ATXN2/rs2073950	A	0.281, 0.239	1.03	0.3102
ATXN2/rs2301621	A	0.283, 0.239	1.127	0.2884
ATXN2/rs10774625	A	0.409, 0.399	0.043	0.8358
ATXN2/**rs10849952**	A	0.952, 0.940	0.308	0.5791
ATXN2/rs6490162	G	0.281, 0.239	1.03	0.3102
ATXN2/rs628825	A	0.319, 0.294	0.327	0.5677
ATXN2/rs630512	A	0.322, 0.294	0.416	0.5187
ATXN2/rs16941541	G	0.948, 0.945	0.018	0.8927
ATXN2/rs7969300	G	0.952, 0.940	0.308	0.5791
ATXN2/rs616513	A	0.322, 0.284	0.738	0.3903
ATXN2/rs12369009	C	0.307, 0.280	0.397	0.5284

Table A1. Most significant ATXN2 haplotype in the UK dataset (χ^2=13.04; adjusted p = 1) and for the international dataset (χ^2=13.21; adjusted p = 1).

	N	Mean Age Of Onset	No hexanucleotide repeat expansion	Hexanucleotide Repeat Expansion
Bulbar	174	63.49	159	15
Limb	364	59.01	343	21
Respiratory	15	66.46	15	-
Mixed	28	57.11	26	2
Total or Mean	581	60.45	543	38

Table A2. Number of cases by limb onset showing AOO and mutation status

Chr	SNP	Base Position	Odds Ratio	Allele	SE	P-Value
9	rs10967976	27534943	1.41	G	0.06349	6.16E-08
9	rs903603	27519316	0.7098	A	0.06375	7.55E-08
9	rs10812611	27532261	1.403	G	0.06347	9.82E-08
6	rs1361121	50032029	0.5254	C	0.1314	9.67E-07
9	rs774359	27551049	1.394	G	0.06842	1.23E-06
9	rs10812605	27500360	1.355	G	0.06437	2.38E-06
9	rs3849942	27533281	1.385	A	0.06996	3.28E-06
11	rs1522659	80518746	0.7222	A	0.07065	4.10E-06
9	rs2814707	27526397	1.375	A	0.06971	5.00E-06
9	rs774357	27549835	1.373	A	0.06983	5.61E-06
8	rs7003470	19386565	1.472	A	0.08526	5.68E-06
4	rs2866197	101325339	0.685	G	0.0834	5.74E-06
2	rs2177083	202669910	0.7547	G	0.06227	6.18E-06
10	rs951030	107579349	1.427	A	0.0794	7.53E-06
10	rs11192617	107572237	1.427	A	0.07949	7.58E-06
12	rs10849295	5543688	0.7077	A	0.07893	1.18E-05
9	rs774352	27506590	1.339	G	0.06707	1.34E-05
9	rs774351	27506640	1.339	C	0.06708	1.36E-05
9	rs1982915	27569560	0.762	A	0.06286	1.53E-05
9	rs7046653	27480967	1.333	A	0.06661	1.56E-05
8	rs1494913	108957826	0.7496	G	0.06675	1.58E-05
4	rs11734827	101327031	0.7002	C	0.08256	1.58E-05
8	rs1389976	108959180	0.7512	G	0.06668	1.79E-05
9	rs10757665	27547919	0.712	G	0.07949	1.92E-05
9	rs12349820	27543876	0.7132	G	0.07947	2.10E-05
5	rs1423515	82098300	1.651	A	0.1192	2.57E-05
4	rs17575890	37005907	1.315	G	0.06532	2.73E-05
19	rs11669124	44315092	0.7529	A	0.06786	2.88E-05
9	rs2073817	135549281	0.7575	A	0.06653	3.00E-05
7	rs2293629	129144534	1.299	G	0.06288	3.11E-05
22	rs138705	37463749	1.318	A	0.06636	3.19E-05
9	rs2453556	27576162	1.298	G	0.06295	3.39E-05
8	rs2941627	106352206	0.642	G	0.1071	3.53E-05
22	rs6004919	24785216	1.525	A	0.1021	3.58E-05
16	rs8053509	8769772	1.418	C	0.8465	3.70E-05
7	rs3736626	129137406	1.296	C	0.6286	3.73E-05
3	rs4684627	9231398	0.7414	G	0.07265	3.80E-05
9	rs129901	135542084	0.7551	G	0.06821	3.83E-05
6	rs9342307	92699953	0.7653	A	0.0651	3.99E-05
9	rs886016	135545450	0.763	G	0.06591	4.07E-05
22	rs6008748	47287746	0.7032	G	0.08603	4.25E-05
11	rs669565	93175487	0.6189	A	0.1174	4.33E-05
21	rs9983113	39237186	1.287	A	0.06182	4.37E-05

7	rs17556430	129133526	1.293	A	0.0629	4.42E-05
5	rs2048213	66312660	1.694	A	0.1291	4.43E-05
2	rs3769185	173721745	0.7704	A	0.06395	4.53E-05
17	rs11655490	68625159	1.337	A	0.07145	4.76E-05
22	rs4823800	47302439	0.6935	C	0.09016	4.92E-05
2	rs17014873	127546689	1.538	C	0.1062	5.01E-05
20	rs6128206	55868747	1.375	G	0.07848	5.03E-05

Table A3. Top 50 SNPs taken from GWAS of All Cases vs. Controls

Chr	SNP	Base Position	Odds Ratio	Allele	SE	P-Value
6	rs1361121	50032029	0.512	C	0.137	1.02E-06
11	rs1522659	80518746	0.7143	A	0.07299	4.03E-06
12	rs10849295	5543688	0.6932	A	0.0819	7.69E-06
4	rs2866197	101325339	0.6847	G	0.08603	1.07E-05
2	rs2177083	202669910	0.7548	G	0.06405	1.12E-05
16	rs8053509	8769772	1.459	C	0.08646	1.24E-05
7	rs10243139	30832290	1.509	G	0.09486	1.46E-05
9	rs903603	27519316	0.7561	A	0.06533	1.87E-05
22	rs6004919	24785216	1.561	A	0.1042	1.91E-05
9	rs2073817	135549281	0.7463	A	0.0687	2.05E-05
5	rs2048213	66312660	1.747	A	0.1313	2.12E-05
4	rs17575890	37005907	1.33	G	0.06713	2.20E-05
4	rs11734827	101327031	0.6972	C	0.08524	2.32E-05
10	rs10996409	66862396	0.6226	G	0.1135	2.99E-05
5	rs1173485	157502407	1.31	G	0.06483	3.18E-05
4	rs2054684	173475020	1.515	A	0.09993	3.24E-05
7	rs2293629	129144534	1.308	G	0.06478	3.35E-05
6	rs9398002	150827248	1.424	G	0.08549	3.50E-05
10	rs951030	107579349	1.404	A	0.08204	3.55E-05
10	rs11192617	107572237	1.404	A	0.08214	3.56E-05
8	rs1494913	108957826	0.7528	G	0.06873	3.61E-05
2	rs17238525	156527831	1.304	G	0.06434	3.62E-05
9	rs10967976	27534943	1.309	G	0.06522	3.63E-05
11	rs10833748	22344321	1.309	A	0.06538	3.87E-05
9	rs10816767	110822491	0.7646	G	0.06524	3.87E-05
7	rs3736626	129137406	1.305	C	0.06477	3.99E-05
8	rs1389976	108959180	0.7544	G	0.06866	4.07E-05
9	rs129901	135542084	0.7494	G	0.07038	4.17E-05
9	rs886016	135545450	0.7577	G	0.06796	4.45E-05
9	rs1537431	110828049	1.3	G	0.06448	4.79E-05
7	rs17556430	129133526	1.301	A	0.0648	4.80E-05
12	rs2970809	4434874	1.333	A	0.07078	4.83E-05
12	rs11051642	31916867	1.355	G	0.07486	4.95E-05
9	rs10812611	27532261	1.303	G	0.0652	5.02E-05
15	rs4932583	90230093	0.7604	G	0.06756	5.04E-05
16	rs11074843	26997812	1.3	A	0.06512	5.53E-05
8	rs7003470	19386565	1.428	A	0.08847	5.60E-05
2	rs10196846	38232614	1.407	A	0.0849	5.69E-05
5	rs1173472	157488401	1.383	A	0.08067	5.76E-05
21	rs9983113	39237186	1.292	A	0.06363	5.78E-05
4	rs7681336	7296855	1.568	A	0.1121	5.98E-05
6	rs9377684	105552057	0.7029	G	0.08788	6.03E-05
14	rs1420813	86546345	1.301	A	0.06559	6.15E-05
2	rs3769185	173721745	0.7681	A	0.0659	6.23E-05
5	rs1614336	157484446	1.297	C	0.06501	6.25E-05

15	rs7175096	90231631	0.723	G	0.08112	6.39E-05
6	rs9384533	150817371	1.415	A	0.08736	7.00E-05
8	rs7827583	128131764	1.403	A	0.08523	7.03E-05
6	rs9342307	92699953	0.7663	A	0.06706	7.20E-05
18	rs1517166	59848277	0.7761	G	0.06392	7.31E-05

Table A4. Top 50 SNPs from GWAS Cases vs. Controls with hexanucleotide mutation cases removed

Chr	SNP	Base Position	Beta	Allele	SE	P-Value
9	rs774357	27549835	5.724	A	0.233	7.67E-85
9	rs2814707	27526397	5.711	A	0.2337	3.02E-84
9	rs3849942	27533281	5.711	A	0.2337	3.02E-84
9	rs774359	27551049	5.496	G	0.2462	1.29E-74
9	rs774352	27506590	5.219	G	0.2604	3.93E-64
9	rs774351	27506640	5.219	C	0.2604	3.93E-64
9	rs7046653	27480967	5.219	A	0.2611	6.86E-64
9	rs10812605	27500360	5.216	G	0.2828	7.93E-57
9	rs10967976	27534943	5.857	G	0.3308	1.72E-53
9	rs10812611	27532261	5.835	G	0.3335	1.50E-52
9	rs10511816	27458461	3.943	A	0.3384	1.44E-27
9	rs2453556	27576162	3.987	G	0.3438	2.35E-27
9	rs4879515	27472235	4.019	A	0.3589	8.33E-26
9	rs903603	27519316	-2.937	A	0.3496	5.98E-16
9	rs2282241	27562255	-2.825	A	0.3568	1.94E-14
9	rs11792285	27509173	-2.491	A	0.3395	1.04E-12
9	rs2282240	27562634	-2.467	A	0.3476	5.00E-12
9	rs1982915	27569560	-2528	A	0.3615	1.00E-11
9	rs10967952	27464214	-2..439	G	0.3589	3.49E-11
9	rs10967959	27472967	-2.544	A	0.3801	6.58E-11
9	rs10967958	27471905	-2.411	A	0.3616	7.77E-11
9	rs12349820	27543876	-2038	G	0.361	2.94E-08
9	rs10757665	27547919	-2.038	G	0.361	2.94E-08

9	rs1977661	27492986	-2.651	A	0.4708	3.20E-08
9	rs1822723	27468052	-1.934	A	0.3472	4.37E-08

Table A5. Top 25 significant SNPs from genome-wide linear regression with repeat number as quantitative variable

Chr	SNP	Base Position	Odds Ratio	Allele	SE	P-Value
9	rs3849942	27533281	4.898	A	0.1144	1.35E-43
9	rs2814707	27526397	4.846	A	0.1139	1.95E-43
9	rs774357	27549835	4.863	A	0.1142	2.03E-43
9	rs774359	27551049	4.786	G	0.1137	6.62E-43
9	rs7046653	27480967	4.09	A	0.1107	6.67E-37
9	rs774352	27506590	4.047	G	0.1103	1.34E-36
9	rs774351	27506640	4.047	C	0.1103	1.35E-36
9	rs10812605	27500360	3.516	G	0.1107	9.75E-30
9	rs10967976	27534943	3.16	G	0.1154	2.69E-23
9	rs10812611	27532261	3.119	G	0.1154	7.88E-23
9	rs903603	27519316	0.3509	A	0.1166	3.28E-19
9	rs10511816	27458461	2.547	A	0.1071	3.08E-18
9	rs4879515	27472235	2.558	A	0.1095	1.16E-17
9	rs2282241	27562255	0.3978	A	0.1198	1.67E-14
9	rs1565948	27549733	0.4259	A	0.1118	2.70E-14
9	rs2453556	27576162	2.096	G	0.1038	1.19E-12
9	rs11792285	27509173	0.4815	A	0.1215	1.97E-09
9	rs1982915	27569560	0.5284	A	0.1068	2.59E-09
9	rs2282240	27562634	0.4178	A	0.1501	6.65E-09
9	rs1822723	27468052	0.4748	A	0.1393	9.70E-08
9	rs10967959	27472967	0.3496	A	0.1987	1.32E-07
9	rs10757665	27547919	0.4618	G	0.1491	2.38E-07
9	rs12349820	27543876	0.4626	G	0.1491	2.53E-07
9	rs10967952	27464214	0.4418	G	0.1656	8.66E-07
18	rs1517166	59848277	0.5999	G	0.1037	8.88E-07

Table A6. Top 25 significant SNPs from GWAS of cases with 7-23 repeats (n=208) vs. controls (n=4142)

Chr	SNP	Base Position	Odds Ratio	Allele	SE	P-Value
9	rs10967976	27534943	2.818	G	0.08569	1.19E-33
9	rs10812611	27532261	2.781	G	0.08562	696E-33
9	rs903603	27519316	0.3854	A	0.08628	2.21E-28
9	rs774359	27551049	2.244	G	0.0817	4.48E-23
9	rs10812605	27500360	2.194	G	0.07978	6.75E-23
9	rs2453556	27576162	2.149	G	0.7923	4.62E-22
9	rs774352	27506590	2.144	G	0.0803	2.10E-21
9	rs774351	27506640	2.144	C	0.08031	2.13E-21
9	rs7046653	27480967	2.123	A	0.08001	5.05E-21
9	rs3849942	27533281	2.175	A	0.08267	5.44E-21
9	rs2814707	27526397	2.157	A	0.08232	1.00E-20
9	rs774357	27549835	2.157	A	0.08245	1.15E-20
9	rs1982915	27569560	0.4897	A	0.0817	2.35E-18
9	rs11792285	27509173	0.4639	A	0.09217	7.81E-17
9	rs4879515	27472235	1.793	A	0.07818	8.10E-14
9	rs2282240	27562634	0.4456	A	0.1097	1.76E-13
9	rs10511816	27458461	1.768	A	0.07776	2.31E-13
9	rs10757665	27547919	0.4827	G	0.1098	3.28E-11
9	rs12349820	27543876	0.4836	G	0.1098	3.63E-11
9	rs2477518	27589746	1.674	G	0.07852	5.37E-11
22	rs138705	37463749	1.525	A	0.07969	1.18E-07
5	rs3806873	5515607	4.413	G	0.3026	9.32E-07
5	rs3806874	5515620	4.328	A	0.3017	1.20E-06
9	rs10967958	27471905	0.5708	A	0.116	1.35E-06
9	rs10122902	27546780	1.523	A	0.08761	1.55E-06

Table A7. Top 25 significant SNPs from GWAS of cases with >2 repeats (n=385) vs. controls (n=4142)

All Regions			Medulla			Cervical			Thoracic			Lumbar		
Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value
SLC1A7	0.27	9.16E-19	SHROOM2	0.60	9.40E-08	ARL6IP1	0.98	4.65E-17	SLC1A7	0.26	1.54E-06	QPRT	1.93	6.82E-08
NCRNA00095	0.31	1.49E-13	NT5C	1.78	7.87E-06	SLC1A7	0.29	9.20E-05	IL7R	3.23	3.51E-05	IL7R	3.00	3.26E-07
MCM3APAS	3.07	1.19E-08	ADH5	1.71	3.73E-05	NCRNA00095	0.31	4.07E-04	SPRY3	2.02	1.48E-03	SLC1A7	0.23	5.05E-06
C21orf62	2.82	4.36E-07	SULT1B1	0.45	3.73E-05	APLN	1.55	1.17E-03	STAT4	1.69	4.52E-03	NT5C	1.80	1.24E-03
CD38	2.08	6.11E-07	SYNC	1.94	7.27E-05	IGFBP3	0.42	2.60E-03	TUFT1	1.63	4.77E-03	FAM162A	1.60	3.59E-03
QPRT	1.80	6.08E-06	SCP2	1.51	4.74E-04	PPEF1	0.74	6.04E-03	CHST9	2.12	6.27E-03	DHFRL1	1.64	4.82E-03
CHST9	1.93	3.53E-05	ZFP36L2	1.60	9.08E-04	PTGS2	0.63	6.27E-03	IGFBP3	0.44	1.16E-02	SLC36A1	1.73	5.25E-03
C8orf34	2.23	2.77E-04	DACT2	0.44	1.21E-03	RASL11A	0.67	1.11E-02	C21orf62	3.71	1.37E-02	RCVRN	0.43	6.10E-03
PJCG6	0.53	3.00E-04	FZD6	0.68	1.21E-03	POTEB	2.08	1.64E-02	GRAMD1C	2.16	1.60E-02	C14orf149	1.76	6.99E-03
PPP1R1A	1.75	5.19E-04	MORC2	0.56	1.21E-03	TNNC1	0.44	2.81E-02	FBXO34	1.52	3.06E-02	ACADSB	1.71	9.70E-03
ACSM5	2.38	9.72E-04	SLC1A7	0.25	1.28E-03	ADA	0.56	2.97E-02	HERC6	1.55	3.06E-02	CYB5R1	1.85	1.11E-02
SHE	0.61	1.16E-03	ZFP1	0.67	1.44E-03	LOC645296	0.62	2.97E-02	LOC595101	0.51	3.29E-02	C17orf91	1.81	1.12E-02
SULT1B1	0.53	1.58E-03	SHE	0.50	1.71E-03				EGFL8	0.43	3.60E-02	GPR27	0.38	1.12E-02
GPR27	0.58	2.24E-03	ANKRD6	1.39	1.87E-03				HIST1H4B	0.44	3.60E-02	NCRNA00095	0.27	1.12E-02
TMEM200A	1.90	3.41E-03	C9orf103	1.51	2.05E-03				MIR555	2.02	3.60E-02	DNAJC27	1.64	1.22E-02
GSTT1	2.24	3.62E-03	EML3	1.38	2.05E-03				SNTB1	1.86	3.79E-02	WDR19	1.49	1.63E-02
LMO2	0.68	3.62E-03	MGEA5	0.63	2.05E-03				TMEM106A	0.43	3.79E-02	MCM3APAS	3.87	2.09E-02
MTBP	1.70	3.62E-03	SERPINF1	1.70	2.75E-03				ADCY8	2.22	3.80E-02	C14orf104	1.69	2.37E-02
ZNF322B	0.46	3.62E-03	LILRB3	1.43	2.78E-03				UFD1L	1.26	3.85E-02	ACPL2	2.08	2.92E-02
KAL1	1.83	4.08E-03	HTT	0.69	2.94E-03				FAM46C	1.69	3.99E-02	PRODH	2.13	3.46E-02
RCVRN	0.46	4.08E-03	LMO2	0.53	2.94E-03							PLEKHG3	1.68	3.53E-02
F5	1.75	4.35E-03	RETSAT	1.52	2.94E-03							SPATA20	1.59	3.65E-02
SNTB1	1.52	4.43E-03	SCUBE2	1.64	2.94E-03							TRUB1	1.75	3.65E-02

FFAR3	0.49	5.40E-03	FBXL17	1.59	3.22E-03				MFN1	1.51	4.04E-02
S100A3	2.01	5.94E-03	RNASET2	1.55	3.22E-03				CYB5D2	1.58	4.35E-02
CCDC102B	1.36	5.99E-03	MARVELD2	0.59	3.40E-03				FRMD4A	1.47	4.35E-02
GALM	1.66	6.06E-03	C1orf115	0.55	4.16E-03				LOC401152	0.57	4.35E-02
SYNC1	1.30	6.06E-03	CLCA2	0.64	4.98E-03				PLEKHH2	1.57	4.35E-02
APLN	1.60	6.41E-03	C12orf5	1.48	5.76E-03				ACAP2	1.45	4.90E-02
DCT	1.67	6.41E-03	GLT25D2	0.69	6.28E-03						
FAM21A	0.43	6.41E-03	DKFZP564J102	1.67	6.72E-03						
IFI27L1	1.30	6.41E-03	ENOX1	0.60	6.94E-03						
LOC731486	2.07	6.41E-03	SIX4	0.52	7.00E-03						
EFCAB2	1.35	6.78E-03	QPRT	1.97	7.13E-03						
ID4	1.31	7.09E-03	INADL	1.64	7.56E-03						
CMTM8	0.74	7.54E-03	RFX4	1.76	7.63E-03						
IGFBP3	0.50	7.65E-03	FAM188A	0.71	7.86E-03						
FMN2	1.45	9.13E-03	CRELD2	0.69	8.15E-03						
KNTC1	1.51	9.35E-03	C21orf56	0.68	8.29E-03						
CD48	2.24	1.00E-02	MSRA	1.33	8.37E-03						
C9orf43	1.58	1.07E-02	ACO2	1.39	8.40E-03						
FHL1	1.38	1.07E-02	PZP	0.53	8.40E-03						
GPR107	0.58	1.07E-02	PFKFB2	1.61	9.40E-03						
LILRA2	1.67	1.07E-02	CMTM8	0.67	9.81E-03						
NT5C	1.59	1.07E-02	STK17B	1.53	9.99E-03						
WDR19	1.34	1.09E-02	CCDC68	0.59	1.03E-02						
IL7R	2.39	1.15E-02	STON1-GTF2A1L	1.35	1.10E-02						
C1orf94	1.83	1.18E-02	MTHFD1L	1.35	1.17E-02						
CMBL	1.42	1.18E-02	KIAA1009	0.69	1.31E-02						
FSBP	1.52	1.18E-02	WDR49	1.34	1.35E-02						

MSRA	1.45	1.18E-02	NOTCH2NL	1.48	1.50E-02
TMEM204	0.63	1.20E-02	ADA	0.53	1.54E-02
GADD45G	1.78	1.23E-02	TARP	0.67	1.57E-02
PFKFB2	1.42	1.25E-02	C20orf191	0.67	1.74E-02
SLC36A1	1.35	1.27E-02	AQP4	1.34	1.87E-02
ADH5	1.51	1.29E-02	HYOU1	0.74	1.89E-02
GRAMD4	1.53	1.31E-02	UBE2Q1	0.68	1.94E-02
CENPK	1.68	1.34E-02	NFATC2	1.73	2.00E-02
CBWD1	1.28	1.36E-02	FZD8	0.67	2.06E-02
RAMP2	0.72	1.36E-02	LRRC25	2.40	2.17E-02
CDKN3	1.60	1.39E-02	PRKCDBP	0.71	2.51E-02
ADAMTS5	1.54	1.44E-02	GPKOW	0.59	2.54E-02
PCDHB16	1.66	1.44E-02	SPATA4	1.55	2.94E-02
LOC595101	0.64	1.46E-02	YJEFN3	0.59	2.94E-02
GPNMB	2.09	1.49E-02	MCM3APAS	3.03	3.01E-02
TFCP2L1	1.72	1.50E-02	PTPRB	0.77	3.01E-02
CENPP	1.53	1.51E-02	MED17	0.67	3.05E-02
LOC645296	0.73	1.51E-02	LOC440396	0.75	3.05E-02
FOXF2	0.68	1.53E-02	ZNF341	0.75	3.05E-02
MSLN	2.28	1.54E-02	GPR62	0.61	3.21E-02
PLCXD3	1.41	1.55E-02	AP2A2	0.48	3.25E-02
C19orf18	1.69	1.57E-02	CCDC102B	1.42	3.25E-02
TSPAN6	1.41	1.61E-02	ACSM5	2.99	3.26E-02
IL18RAP	1.62	1.68E-02	FAM149A	1.45	3.26E-02
TDRD9	1.89	1.69E-02	GNG12	1.43	3.26E-02
C12orf73	0.65	1.78E-02	RAB12	1.64	3.26E-02
DNASE2	1.42	1.85E-02	RNASEH1	0.68	3.26E-02
THSD1	0.64	1.91E-02	TMEM204	0.63	3.26E-02

MARVELD2	0.66	1.93E-02	GEM	1.92	3.30E-02
TAS2R43	0.52	1.96E-02	RREB1	1.44	3.35E-02
FUT9	1.40	2.04E-02	PLTP	2.02	3.36E-02
SNORD54	1.66	2.04E-02	LILRA2	2.35	3.39E-02
PALB2	1.36	2.07E-02	PDIA4	0.64	3.39E-02
PROK2	1.94	2.07E-02	RPS6KA1	1.56	3.48E-02
SEPT3	1.30	2.09E-02	LCA5L	1.71	3.54E-02
ELTD1	0.75	2.09E-02	LOC54103	1.49	3.61E-02
MAMLD1	1.41	2.09E-02	IGFBP6	0.71	3.63E-02
SLC41A1	1.37	2.16E-02	TTC32	1.39	3.67E-02
GAN	1.43	2.24E-02	MPP6	1.49	3.75E-02
C4orf19	1.78	2.25E-02	C20orf54	0.71	3.77E-02
DCLRE1B	1.49	2.25E-02	HAPLN2	0.61	3.77E-02
SAP30BP	0.73	2.25E-02	SMARCAD1	0.71	3.77E-02
HS2ST1	1.30	2.28E-02	PODN	0.68	3.81E-02
FLJ10357	1.34	2.40E-02	JUNB	1.83	3.89E-02
DDIT4L	1.42	2.42E-02	GPR85	0.68	3.91E-02
ADCY8	1.51	2.43E-02	FANCE	0.76	3.95E-02
S100A4	1.44	2.51E-02	PUS7	0.76	4.08E-02
FLJ22536	1.34	2.52E-02	LOC113386	0.64	4.10E-02
EAF2	1.29	2.53E-02	INTS7	0.74	4.11E-02
HLA-A29.1	0.49	2.59E-02	CRTAP	1.37	4.27E-02
CDC42EP4	1.27	2.69E-02	CBX5	1.28	4.37E-02
MTFR1	1.43	2.72E-02	BTNL9	0.69	4.39E-02
FBXL17	1.39	2.72E-02	SLC22A23	0.76	4.39E-02
PPP4R4	1.47	2.73E-02	TAS2R46	0.67	4.50E-02
ZNF323	0.72	2.75E-02	CLPTM1L	0.60	4.51E-02
PLK4	1.51	2.75E-02	CD38	2.01	4.70E-02

CHRNA1	4.32	2.76E-02	ZIC4	0.61	4.72E-02
C1orf110	1.51	2.77E-02	RAMP2	0.64	4.76E-02
LOC644928	0.61	2.84E-02	MEF2C	1.57	4.77E-02
ANO6	1.32	2.90E-02	SFMBT2	1.52	4.81E-02
RPESP	1.75	2.90E-02	CEBPA	1.76	4.83E-02
SERPINE1	1.88	2.90E-02	NEUROD2	0.44	4.83E-02
FDX1	1.37	2.94E-02	TAS2R43	0.43	4.83E-02
TRIM22	1.52	2.94E-02	ANKAR	0.79	4.92E-02
CPA4	0.64	2.94E-02	FLJ13614	1.50	4.99E-02
NDUFA7	1.74	3.00E-02			
DPEP3	2.02	3.06E-02			
C9orf103	1.44	3.08E-02			
C2	1.59	3.08E-02			
C4orf31	1.54	3.09E-02			
C4orf16	1.38	3.14E-02			
C2orf40	1.57	3.15E-02			
MMRN1	0.64	3.15E-02			
PCDHB7	1.44	3.16E-02			
C6orf192	1.25	3.18E-02			
SLC15A3	1.53	3.18E-02			
KLRB1	1.90	3.34E-02			
PLCD3	1.43	3.36E-02			
SLC15A2	1.36	3.49E-02			
OSBPL3	1.53	3.50E-02			
COBL	0.75	3.52E-02			
SSPN	1.23	3.62E-02			
MS4A14	1.90	3.69E-02			
GZMK	1.97	3.74E-02			

RPL29	0.70	3.76E-02
DDX49	0.68	3.79E-02
C14orf149	1.35	3.91E-02
WASF2	0.79	4.01E-02
C3orf51	0.66	4.02E-02
LOC283932	0.76	4.06E-02
LRRC25	1.96	4.07E-02
PAPSS2	0.68	4.08E-02
KIAA0182	1.39	4.24E-02
ZNF7	1.77	4.28E-02
PRODH	1.67	4.36E-02
C8orf40	1.40	4.47E-02
CDNF	1.62	4.47E-02
CYB5R1	1.49	4.47E-02
SMAD6	0.73	4.47E-02
USP37	0.78	4.47E-02
ZDHHC1	1.54	4.47E-02
C11orf92	1.48	4.58E-02
CHSY3	1.52	4.68E-02
GJA1	1.25	4.68E-02
RPE65	2.02	4.69E-02
HERC2P2	0.71	4.71E-02
LOC442582	0.67	4.71E-02
BTNL9	0.71	4.71E-02
IGSF5	1.51	4.78E-02
MIR575	0.72	4.78E-02
TUBGCP3	1.45	4.78E-02
ABCC9	1.34	4.86E-02

FMO4	1.54	4.90E-02				
IL3RA	0.58	4.92E-02				
ANKRD13B	0.73	4.93E-02				
FLJ33360	0.70	4.96E-02				
NOL12	0.78	4.96E-02				

Table A8. All significantly differentially expressing genes by anatomical region

Glycosylation & Transmembrane activity			Symporter activity			Glycan metabolic processes			Calcium Binding		
Gene	Fold Change	Difference P-value	Gene	Fold Change	Difference P-value	Gene	Fold Change	Difference P-value	Gene	Fold Change	Difference P-value
SLC36A1	1.35	0.01	SLC36A1	1.35	0.01	C4ORF31	1.54	0.03	S100A4	1.44	0.03
CYB5R1	1.49	0.04	SLC15A2	1.36	0.03	MSRA	1.45	0.01	S100A3	2.01	0.01
SLC15A2	1.36	0.03	SLC1A7	0.27	0.00	CHST9	1.93	0.00	PCDHB7	1.44	0.03
ADCY8	1.51	0.02	SLC15A3	1.53	0.03	GSTT1	2.24	0.00	F5	1.75	0.00
WASF2	0.79	0.04				PAPSS2	0.68	0.04	PCDHB16	1.66	0.01
TSPAN6	1.41	0.02				HS2ST1	1.30	0.02	SNTB1	1.52	0.00
GJA1	1.25	0.05							PLCD3	1.43	0.03
RPE65	2.02	0.05							EFCAB2	1.35	0.01
HS2ST1	1.30	0.02							ELTD1	0.75	0.02
C8ORF40	1.40	0.04							RCVRN	0.46	0.00
CD48	2.24	0.01							ANO6	1.32	0.03
LILRA2	1.67	0.01									
SLC1A7	0.27	0.00									
DPEP3	2.02	0.03									
ELTD1	0.75	0.02									
CHRNA1	4.32	0.03									
ANO6	1.32	0.03									
KLRB1	1.90	0.03									
RAMP2	0.72	0.01									
ZDHHC1	1.54	0.04									
PCDHB7	1.44	0.03									

TMEM204	0.63	0.01
BTNL9	0.71	0.05
IL18RAP	1.62	0.02
TMEM200A	1.90	0.00
SSPN	1.23	0.04
LRRC25	1.96	0.04
IGSF5	1.51	0.05
CD38	2.08	0.00
CHSY3	1.52	0.05
CHST9	1.93	0.00
SLC41A1	1.37	0.02
C6ORF192	1.25	0.03
FUT9	1.40	0.02
FFAR3	0.49	0.01
IL7R	2.39	0.01
GRAMD4	1.53	0.01
FMO4	1.54	0.05
DCT	1.67	0.01
MARVELD2	0.66	0.02
GPR27	0.58	0.00
PCDHB16	1.66	0.01
MSLN	2.28	0.02
TAS2R43	0.52	0.02
USP37	0.78	0.04
GPNMB	2.09	0.01
GPR107	0.58	0.01
IFI27L1	1.30	0.01
ABCC9	1.34	0.05

THSD1	0.64	0.02			
CMTM8	0.74	0.01			
C19ORF18	1.69	0.02			
SLC15A3	1.53	0.03			
C11ORF92	1.48	0.05			
MS4A14	1.90	0.04			
IL3RA	0.58	0.05			

Table A9. Differentially expressing genes by functional cluster comparing cases and controls

Regulation of insulin, peptide and hormone secretion		
Gene	**Fold Change**	**Difference P-Value**
CD38	2.01	0.05
PFKFB2	1.61	0.01
MGEA5	0.63	0.00
ADH5	1.71	0.00
GNG12	1.43	0.03
JUNB	1.83	0.04

Table A10. Differentially expressed genes by functional cluster comparing cases and controls medulla samples

Regulation of muscle contraction, response to external stimulus & apoptosis		
Gene	**Fold Change**	**Difference P-Value**
PTGS2	0.63	0.01
PPEF1	0.74	0.01
TNNC1	0.44	0.03
SLC1A7	0.29	0.00
POTEB	2.08	0.02
IGFBP3	0.42	0.00
ADA	0.56	0.03

Table A11. Differentially expressed genes by functional cluster comparing cases and controls cervical samples

Conjugation pathway & catabolic processes		
Genes	**Fold Change**	**Difference P-value**
UFD1L	1.26	0.04
HERC6	1.55	0.03
FBXO34	1.52	0.03

Table A12. Differentially expressed genes by functional cluster comparing cases and controls thoracic samples

Flavoproteins and oxidoreductase			Pleckstrin homology		
Gene	**Fold Change**	**Differential P-Value**	**Gene**	**Fold Change**	**Differential P-Value**
ACADSB	1.71	0.01	PLEKHG3	1.68	0.04
CYB5R1	1.85	0.01	PLEKHH2	1.57	0.04
DHFRL1	1.64	0.00	FRMD4A	1.47	0.04
PRODH	2.13	0.03	ACAP2	1.45	0.05

Table A13. Differentially expressed genes by functional cluster comparing cases and controls lumbar samples

Onset Region - Cervical			Second Region - Thoracic			Third Region - Lumbar			Fourth Region - Medulla		
Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value
SLC1A7	0.20	6.15E-04	SLC1A7	0.22	1.90E-06	SLC1A7	0.15	2.89E-10	FAM43A	0.55	2.11E-04
POTEB	2.26	1.12E-02	IL7R	2.92	7.30E-04	IL7R	3.17	3.11E-07	MCM3APAS	3.34	1.06E-03
						QPRT	1.94	4.68E-06	SHROOM2	0.63	9.60E-03
						ACSM5	2.56	4.28E-05	SOX17	0.60	9.60E-03
						NT5C	1.92	1.50E-04	SULT1B1	0.46	1.34E-02
						ACER2	0.44	1.14E-02	CENPV	0.60	3.26E-02
						DHFRL1	1.68	1.40E-02			
						C14orf104	1.75	4.53E-02			
						RCVRN	0.47	4.69E-02			

Table A14. All significantly differentially expressing genes by disease spread

Mild-Moderate			Moderate			Moderate-Severe			Severe		
Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value	Gene Symbol	Fold Change	P-Value
NCRNA00095	0.19	2.16E-33	SLC1A7	0.26	2.96E-05	SLC1A7	0.15	2.16E-33	ANGPT2	0.36	1.55E-13
FAM21A	0.08	9.50E-19	LGI4	0.58	1.04E-04	PJCG6	0.44	2.85E-06	ELSPBP1	3.44	2.43E-08
HLA-DRB4	0.09	2.56E-15	HLA-A29.1	0.17	3.36E-04	MTBP	1.92	7.74E-06	INPP4B	1.68	2.50E-07

MCM3APAS	3.54	8.81E-11	GEMIN5	0.62	2.05E-03	HAPLN3	0.46	6.73E-05	PJCG6	0.43	4.87E-07
SOX18	0.41	4.16E-10	CPA4	0.45	2.53E-03	ADA	0.50	7.24E-04	TKTL1	2.03	1.53E-06
CH25H	2.47	4.85E-08	TLE2	0.50	2.53E-03	RCVRN	0.41	8.77E-04	HIVEP3	0.51	4.18E-06
CMTM8	0.58	6.07E-08	LOC440248	0.38	3.57E-03	A2LD1	1.70	8.77E-04	MCM3APAS	4.15	5.41E-06
QPRT	2.02	4.82E-07	NCRNA00095	0.37	7.31E-03	CCDC102B	1.43	1.14E-03	SLC1A7	0.34	6.19E-06
ADCY8	2.02	1.34E-06	GPR107	0.46	7.31E-03	TMEM204	0.57	1.20E-03	QDPR	0.46	1.75E-05
TBC1D2B	1.85	4.98E-06	BTN3A2	1.64	7.31E-03	MSRA	1.56	1.38E-03	GALNT6	0.25	2.35E-05
DACT2	0.31	6.25E-06	SULT1B1	0.45	7.35E-03	BTNL9	0.59	1.52E-03	ADA	0.43	3.12E-05
CCDC68	0.38	7.52E-06	TMEM139	0.50	7.35E-03	C9orf103	1.56	2.57E-03	PTGS2	0.40	3.14E-05
CD38	2.39	8.55E-06	GPR27	0.47	8.46E-03	NT5C	1.63	3.72E-03	RCVRN	0.36	4.87E-05
ROBO4	0.36	1.01E-05	ZNF385C	0.47	1.13E-02	ALPL	0.72	5.97E-03	CNKSR1	0.48	9.21E-05
RAMP2	0.54	1.51E-05	RP5-1022P6.2	0.68	1.17E-02	C4orf33	1.41	7.66E-03	LOC645296	0.57	1.25E-04
KAL1	2.53	1.94E-05	RAMP2	0.65	1.28E-02	THSD1	0.59	7.82E-03	ZNF135	1.64	3.24E-04
TNFSF10	0.35	2.92E-05	LOC442582	0.52	1.36E-02	GPNMB	2.50	8.13E-03	NOS1AP	0.62	3.31E-04
IRF5	2.56	3.09E-05	C1orf110	1.71	1.53E-02	C10orf107	1.58	9.26E-03	ZSCAN12L1	0.56	3.31E-04
PIM1	2.07	3.73E-05	PRPF6	0.61	1.67E-02	LILRA2	1.77	9.26E-03	RPE65	2.64	3.83E-04
SLC25A34	0.53	4.65E-05	SHE	0.64	2.17E-02	DCHS1	0.63	1.10E-02	FLJ16165	0.39	4.87E-04
SULT1B1	0.37	5.42E-05	CFH	0.57	2.32E-02	TOMM20L	1.58	1.10E-02	QPRT	1.93	5.42E-04
AEBP1	1.87	7.70E-05	DDIT4L	1.60	2.54E-02	QPRT	1.79	1.10E-02	IGSF1	1.84	9.51E-04
CHST9	2.38	8.23E-05	FFAR3	0.48	2.96E-02	ANGPT2	0.50	1.10E-02	SCUBE2	1.64	1.04E-03
GPR27	0.54	1.01E-04	WIZ	0.54	2.96E-02	SLC19A3	0.67	1.16E-02	PLCD3	1.59	1.08E-03
PRRX1	1.64	1.17E-04	S100P	0.42	2.96E-02	CD38	2.17	1.16E-02	SHROOM4	0.25	1.12E-03
DNASE2	1.69	1.61E-04	ODF3L2	0.60	3.16E-02	SLC5A6	0.73	1.30E-02	PUS7	0.69	1.16E-03
DDIT4L	1.79	1.61E-04	ADCY4	0.59	3.18E-02	NOL12	0.68	1.31E-02	ACSM5	2.56	1.18E-03
LOC645296	0.57	1.70E-04	HLA-C	2.51	3.18E-02	CEBPA	1.50	1.35E-02	CYP3A5	1.74	1.34E-03
TMEM2	0.59	2.02E-04	PADI4	0.48	3.64E-02	FFAR3	0.52	1.38E-02	ADAL	0.59	1.67E-03
C19orf18	2.01	2.11E-04	BGLAP	0.60	3.90E-02	LYRM2	1.58	1.38E-02	CCDC33	2.76	1.70E-03

EMX2OS	0.34	2.13E-04	MIR575	0.67	4.88E-02	FSBP	1.60	1.38E-02	FBXL16	0.46	1.73E-03
GPR116	0.65	2.15E-04	HLA-H	0.60	4.88E-02	SERPINE1	2.06	1.53E-02	C1orf110	1.94	1.91E-03
ABO	0.40	2.77E-04	RCVRN	0.49	4.97E-02	C19orf18	1.82	1.65E-02	ZNF323	0.67	2.05E-03
PIRT	1.62	2.84E-04	LOC595101	0.62	4.97E-02	HLA-A29.1	0.21	1.67E-02	FAM21A	0.38	2.24E-03
HSPA12B	0.56	2.98E-04				HERC2P2	0.64	1.67E-02	DCT	1.92	2.62E-03
GPR124	0.53	3.16E-04				SNORD54	1.65	1.67E-02	GADD45G	2.22	2.83E-03
PTPRB	0.67	3.80E-04				PCDHGB5	0.71	1.68E-02	YPEL3	0.47	2.83E-03
GSTT1	2.28	3.92E-04				OSGIN2	1.35	1.68E-02	TCTN2	1.62	2.83E-03
ABCC3	3.11	4.39E-04				MSLN	2.30	1.68E-02	DNASE2B	2.00	2.90E-03
NRG3	1.68	5.13E-04				C14orf149	1.44	1.69E-02	PCDHB16	1.91	3.01E-03
SHE	0.51	6.64E-04				APLN	1.82	1.74E-02	LOC441743	0.65	3.01E-03
CMBL	1.62	6.91E-04				ABHD7	0.71	1.76E-02	SPATA5L1	0.69	3.50E-03
IL17RB	1.80	7.12E-04				ZIC3	0.55	1.84E-02	ZNF642	0.57	3.51E-03
OXTR	0.63	7.65E-04				PFKFB2	1.44	2.04E-02	CDC42EP4	1.38	3.55E-03
LMO2	0.57	7.75E-04				DACT2	0.52	2.12E-02	WDR27	1.57	3.55E-03
DKFZp451A211	0.33	7.75E-04				ZNF322B	0.39	2.16E-02	SLC16A9	1.38	3.55E-03
SHROOM4	0.23	8.43E-04				PDE10A	0.54	2.16E-02	DNAH2	2.47	3.55E-03
PRODH	2.42	8.57E-04				SHE	0.61	2.16E-02	CHRNA1	2.90	3.56E-03
RGMA	1.72	9.04E-04				EPB42	0.55	2.17E-02	MASP1	1.85	3.99E-03
SLC1A7	0.36	9.20E-04				FUT9	1.38	2.17E-02	FFAR3	0.48	4.00E-03
CHST6	1.56	9.20E-04				ACADSB	1.45	2.20E-02	SPOCK3	0.47	4.00E-03
C1QTNF5	1.61	9.20E-04				LOC54103	1.52	2.29E-02	KLHL2	0.72	4.00E-03
ANKRD38	0.62	9.56E-04				CD34	0.73	2.32E-02	ZNF785	0.60	4.16E-03
CYYR1	0.68	1.14E-03				LOC728323	1.63	2.39E-02	GSTM5	1.81	4.16E-03
PLA1A	0.63	1.19E-03				LMO2	0.66	2.50E-02	GSTM2	1.65	4.27E-03
LOC731486	2.41	1.37E-03				CMTM1	1.29	2.51E-02	C17orf75	1.41	4.33E-03
IGDCC4	1.67	1.42E-03				ADH5	1.61	2.53E-02	LOC644928	0.55	4.70E-03
C20orf160	0.61	1.44E-03				VEGFC	0.54	2.54E-02	ZUFSP	0.70	5.12E-03

BEGAIN	0.46	1.48E-03
LOC401233	2.48	1.48E-03
C6orf48	0.74	1.49E-03
NTNG1	1.53	1.50E-03
FOXQ1	0.69	1.69E-03
SNX24	0.67	1.69E-03
AFAP1L2	0.61	1.71E-03
GRAMD4	1.71	1.73E-03
TRIL	1.48	1.73E-03
C21orf62	3.74	1.99E-03
KCNA5	1.67	2.19E-03
GRASP	0.47	2.22E-03
CABC1	1.72	2.24E-03
TDRD10	0.36	2.30E-03
ABCG2	0.61	2.33E-03
PAX6	1.82	2.39E-03
JAG1	0.65	2.55E-03
F5	2.35	2.65E-03
NAPB	3.26	2.69E-03
BMP8B	0.43	2.74E-03
TDH	1.98	2.77E-03
RPESP	1.83	2.81E-03
NKD1	0.65	2.86E-03
ST14	2.34	2.91E-03
UBQLNL	0.43	2.99E-03
MTBP	1.84	3.11E-03
KNTC1	1.65	3.15E-03
ABCB1	0.69	3.26E-03
C9orf43	1.73	2.54E-02
AHSP	0.53	2.56E-02
ZWINT	1.51	2.61E-02
KANK3	0.59	2.63E-02
IL27RA	0.65	2.63E-02
CA1	0.30	2.67E-02
LOC731486	2.23	2.67E-02
CMTM8	0.71	2.71E-02
RAB11FIP4	0.63	2.76E-02
PLK4	1.62	2.76E-02
C21orf62	2.67	2.76E-02
MTHFD2L	0.74	2.78E-02
FOXF2	0.63	2.82E-02
C1orf94	2.02	2.90E-02
POFUT2	1.03	2.90E-02
FBXL17	1.42	2.91E-02
RAMP2	0.74	2.91E-02
GAN	1.51	2.91E-02
RHOXF1	0.45	2.93E-02
EMR4	2.17	3.12E-02
ACER2	0.57	3.17E-02
IL3RA	0.54	3.21E-02
LRP8	0.76	3.23E-02
PXMP3	1.34	3.45E-02
DHDH	1.37	3.45E-02
MRS2	1.54	3.48E-02
FBXO36	1.38	3.49E-02
DCT	1.82	3.50E-02
STON2	1.54	5.17E-03
C7orf50	1.36	5.34E-03
SNORD49A	0.66	5.51E-03
OR4F21	1.63	5.51E-03
NT5C	1.61	6.42E-03
SNTB1	1.52	6.55E-03
TIPIN	1.40	6.62E-03
ENTPD2	2.15	6.64E-03
METTL8	0.69	6.68E-03
ZNF441	1.64	6.70E-03
LIAS	0.72	6.80E-03
GRAMD4	1.59	6.91E-03
CCDC37	2.17	7.23E-03
C16orf62	0.66	7.88E-03
C14orf79	1.66	8.04E-03
CHSY3	1.48	8.34E-03
POLR2D	0.74	8.50E-03
SNORD54	1.62	8.52E-03
IRX3	1.47	8.52E-03
C3orf51	0.54	9.09E-03
LRAT	1.38	9.09E-03
SNTA1	1.45	9.09E-03
ACSL1	0.67	9.33E-03
GIMAP7	0.71	9.89E-03
IL7R	2.25	9.95E-03
GZMK	1.81	9.96E-03
CES7	0.36	9.96E-03
RHBDL1	0.48	9.97E-03

GJA1	1.44	3.31E-03
HERC2P4	0.59	3.36E-03
MYO1F	1.94	3.46E-03
TBX2	0.56	3.56E-03
TC2N	1.86	3.67E-03
ZNF771	0.55	3.67E-03
ANKRD13B	0.51	3.76E-03
MARVELD2	0.48	3.79E-03
LEPR	0.57	3.87E-03
PAFAH1B3	1.54	4.29E-03
BTNL9	0.66	4.30E-03
FCGRT	1.59	5.23E-03
VIL2	1.41	5.26E-03
NAGLU	1.73	5.26E-03
LOC440563	0.45	5.35E-03
TRAPPC2P1	1.63	5.41E-03
ADCY4	0.65	5.41E-03
PLCXD3	1.80	5.85E-03
PPP4R4	1.80	6.44E-03
HRK	0.42	6.56E-03
IFI27L1	1.42	6.75E-03
CHPF	1.86	6.78E-03
LOC440248	0.42	6.78E-03
DCLK2	1.88	6.99E-03
CXCL14	0.53	7.39E-03
MIF	1.57	7.39E-03
DEGS2	0.58	7.67E-03
TSSC1	0.68	8.18E-03
ECSIT	0.63	3.50E-02
PRELID1	0.65	3.50E-02
GGT5	0.71	3.51E-02
EVI1	0.72	3.53E-02
PLCD3	1.57	3.56E-02
DEGS2	0.62	3.63E-02
DDX49	0.66	3.63E-02
FMO4	1.65	3.63E-02
MARVELD2	0.62	3.69E-02
VTA1	1.50	3.83E-02
PDIA4	0.71	3.86E-02
ZNF7	1.84	3.89E-02
MFSD2	0.65	3.90E-02
ADAMTS5	1.51	3.90E-02
GSTT1	2.55	3.90E-02
KNTC1	1.56	3.96E-02
ABCG2	0.73	4.00E-02
HBD	0.27	4.02E-02
PCDHB16	1.63	4.02E-02
TARP	0.69	4.13E-02
FOXQ1	0.69	4.13E-02
KCND1	1.55	4.13E-02
BRCC3	1.59	4.13E-02
TMEM55A	1.60	4.13E-02
TTC32	1.59	4.17E-02
TPH1	1.59	4.18E-02
C1orf69	0.62	4.20E-02
PAPSS2	0.61	4.21E-02
TUBAL3	0.42	1.10E-02
ECT2	0.72	1.12E-02
DBI	1.54	1.13E-02
BRF2	0.65	1.13E-02
NCRNA00095	0.32	1.14E-02
ADRA1A	1.73	1.16E-02
COL5A2	1.46	1.21E-02
FMO4	1.82	1.25E-02
USMG5	0.36	1.25E-02
SNORD94	4.59	1.25E-02
SERPINE1	1.73	1.28E-02
RPL29	0.65	1.32E-02
KCND1	1.77	1.32E-02
PCSK7	0.69	1.32E-02
ABHD12B	0.58	1.33E-02
STC1	0.54	1.34E-02
CENPN	0.70	1.38E-02
TXNDC3	0.58	1.39E-02
LOC442245	2.12	1.48E-02
LRRC58	0.72	1.50E-02
C2orf70	1.74	1.55E-02
CACNA1B	0.50	1.55E-02
SNAPC3	0.50	1.66E-02
DKFZp434I1020	1.68	1.66E-02
HSD11B1	0.53	1.70E-02
HMGCR	0.66	1.72E-02
WNK2	1.71	1.75E-02
ATP13A4	1.44	1.76E-02

SLC41A1	1.56	8.55E-03
IGFBP3	0.46	8.57E-03
SNTB1	1.58	9.69E-03
REM1	0.51	9.69E-03
TOMM34	0.67	1.01E-02
PLA2G15	1.53	1.01E-02
PIPOX	1.57	1.03E-02
C6orf192	1.40	1.06E-02
C4orf31	1.93	1.09E-02
ID4	1.42	1.15E-02
HLA-DRB1	2.72	1.17E-02
F3	1.67	1.17E-02
LOC401152	0.44	1.17E-02
CFH	0.63	1.19E-02
OLFML2A	0.62	1.20E-02
CENPK	1.95	1.22E-02
OTOS	3.44	1.22E-02
GIMAP7	0.75	1.24E-02
CD44	1.23	1.24E-02
ZNF785	0.53	1.25E-02
TEK	0.69	1.25E-02
HBQ1	0.44	1.25E-02
CLIC5	0.42	1.25E-02
C9orf61	1.76	1.25E-02
STXBP2	1.96	1.25E-02
ZNF507	0.52	1.26E-02
TMEM204	0.47	1.29E-02
EDN1	0.56	1.30E-02
ZFHX2	0.63	4.21E-02
INO80B	1.43	4.21E-02
S100A3	1.94	4.30E-02
PKN3	0.61	4.33E-02
TDRD10	0.48	4.39E-02
IFRD2	0.72	4.39E-02
SIAH2	0.74	4.40E-02
VWA2	0.63	4.47E-02
DBI	1.44	4.64E-02
CBWD1	1.30	4.70E-02
ARHGEF15	0.64	4.79E-02
ALDH7A1	1.36	4.79E-02
FAM131C	2.06	4.80E-02
MORC2	0.66	4.81E-02
KIF26A	0.58	4.84E-02
EGFL7	0.63	4.86E-02
WDR19	1.35	4.89E-02
UBE2QP2	0.74	4.91E-02
C12orf5	1.43	4.91E-02
GORASP1	1.44	4.91E-02
GRAMD4	1.52	4.91E-02
HFM1	1.56	4.91E-02
DENND1C	1.62	4.91E-02
HMG20A	1.39	4.97E-02
ZNF519	1.35	1.88E-02
ACACB	1.39	1.91E-02
MICAL2	0.58	1.93E-02
SSH2	0.62	1.93E-02
APLN	1.61	1.94E-02
CLASP1	0.68	1.94E-02
FOSL1	2.10	1.94E-02
GBP5	1.60	1.94E-02
KIAA0773	0.71	1.97E-02
ZNF7	1.85	2.01E-02
FUT11	0.67	2.01E-02
DKFZP564O0523	0.74	2.01E-02
IGFBP3	0.38	2.02E-02
CAMK2G	1.35	2.04E-02
ZNF322B	0.28	2.07E-02
ZNF284	2.05	2.17E-02
SFTPD	1.85	2.17E-02
CA1	0.32	2.22E-02
CYP4F3	1.98	2.23E-02
ALDH5A1	1.39	2.35E-02
VPS37D	0.65	2.35E-02
CUTA	0.93	2.35E-02
AGFG2	1.52	2.36E-02
SMA5	0.89	2.42E-02
TASP1	1.39	2.42E-02
TTC21A	1.40	2.61E-02
WDR31	1.53	2.71E-02
LRFN1	0.56	2.76E-02

DCT	1.77	1.31E-02
SELP	0.38	1.33E-02
MS4A14	2.22	1.33E-02
APLN	1.68	1.42E-02
CENPP	1.85	1.42E-02
NUDT16	0.73	1.42E-02
MIR575	0.65	1.43E-02
YDJC	1.65	1.43E-02
ITGB2	1.48	1.44E-02
PFKFB2	1.52	1.47E-02
C9orf103	1.54	1.49E-02
FOXF2	0.60	1.49E-02
SIRPB1	1.83	1.49E-02
ACVRL1	0.63	1.51E-02
IL6	2.04	1.51E-02
LOC100129550	1.66	1.57E-02
ELTD1	0.64	1.57E-02
USP37	0.72	1.61E-02
37865	1.45	1.62E-02
LILRA2	2.00	1.65E-02
LOC285016	1.85	1.69E-02
EDN3	0.48	1.75E-02
SUSD1	1.55	1.76E-02
FMN2	1.56	1.81E-02
DNAJC1	1.50	1.81E-02
SLC4A11	1.89	1.82E-02
MALAT1	0.48	1.83E-02
NAALADL1	0.62	1.83E-02
LRFN4	0.62	2.77E-02
LOC220594	0.60	2.80E-02
ZMYND19	0.76	2.86E-02
SVOPL	2.07	2.88E-02
GIMAP8	0.71	2.89E-02
PPP1R3C	0.76	2.89E-02
C1orf158	2.30	2.90E-02
SIX4	0.47	3.02E-02
COBL	0.73	3.02E-02
HERC6	1.37	3.02E-02
GATS	0.77	3.02E-02
CELSR3	0.70	3.02E-02
RAB17	0.65	3.02E-02
AMOT	1.40	3.06E-02
SYNC1	1.33	3.14E-02
CRYGS	0.50	3.14E-02
AGAP11	1.60	3.14E-02
GPR3	0.46	3.17E-02
ASXL2	0.70	3.21E-02
PTPN20A	0.39	3.21E-02
FLJ42957	1.44	3.23E-02
FNBP1L	1.31	3.25E-02
ARPC5L	0.74	3.25E-02
CCBP2	1.56	3.32E-02
ZDHHC1	1.55	3.36E-02
NDUFB1	0.60	3.37E-02
C7orf10	1.37	3.37E-02
LCA5L	1.98	3.44E-02

SH3PXD2A	0.74	1.83E-02
TRIM47	1.45	1.86E-02
MTFR1	1.55	1.88E-02
CNTNAP1	1.52	1.99E-02
MS4A7	1.85	2.00E-02
AOAH	1.75	2.02E-02
EFCAB3	2.21	2.09E-02
TCTEX1D2	1.37	2.11E-02
FOXC1	0.67	2.11E-02
CCL2	1.63	2.12E-02
SERPINA3	1.50	2.12E-02
PTPRO	1.78	2.28E-02
DGKE	0.66	2.29E-02
C4orf19	2.01	2.29E-02
TCF15	0.50	2.29E-02
C12orf39	1.77	2.35E-02
C6orf114	0.67	2.35E-02
LOC441743	0.68	2.39E-02
ZSCAN23	0.52	2.39E-02
C2orf40	1.66	2.42E-02
POMC	0.58	2.42E-02
S100A4	1.53	2.43E-02
NET1	0.79	2.45E-02
PLCD3	1.49	2.49E-02
HBA1	0.36	2.49E-02
MIR320B2	0.55	2.49E-02
AGXT2L1	1.38	2.50E-02
TACSTD1	0.55	2.61E-02
WDSOF1	0.69	3.45E-02
IFT122	1.39	3.53E-02
LOC731486	2.02	3.60E-02
H1F0	1.38	3.61E-02
SGTB	1.90	3.61E-02
KPNA1	0.75	3.66E-02
C12orf39	1.75	3.68E-02
PROK2	3.10	3.74E-02
PLK5P	0.61	3.74E-02
NAPB	2.34	3.78E-02
PLCB3	1.43	3.78E-02
CTDSPL	1.40	3.78E-02
PPP1R13L	1.34	3.79E-02
MYOZ1	1.35	3.82E-02
CALR	0.65	3.83E-02
LOC440563	0.61	3.84E-02
NOTCH2NL	1.44	3.85E-02
PDE10A	0.57	3.93E-02
SLC25A19	0.71	3.94E-02
C15orf63	0.62	3.94E-02
C15orf52	1.33	4.02E-02
GNAO1	1.05	4.13E-02
KIAA1324L	0.77	4.17E-02
SETD7	1.37	4.21E-02
LOC162632	0.74	4.23E-02
FLJ10357	1.45	4.26E-02
MRPL34	0.69	4.39E-02
C11orf92	1.81	4.39E-02

FGF11	1.38	2.67E-02
AGFG2	1.64	2.67E-02
ARSG	0.60	2.73E-02
RAB40AL	0.65	2.73E-02
ZBTB7B	1.76	2.73E-02
FGD5	0.57	2.73E-02
KIAA1009	0.52	2.73E-02
PTPRZ1	1.49	2.73E-02
GAN	1.55	2.80E-02
VANGL1	0.46	2.82E-02
EDC4	0.60	2.82E-02
ZNF280D	0.62	2.84E-02
TMEM45B	0.56	2.91E-02
PCDHGB5	0.71	2.93E-02
FHL1	1.39	2.97E-02
AFG3L1	0.79	2.97E-02
MOAP1	1.41	2.98E-02
EMCN	0.62	2.99E-02
WDR19	1.39	2.99E-02
EI24	0.71	3.00E-02
CDCA7	2.23	3.00E-02
CYP2S1	1.98	3.04E-02
HYI	1.53	3.07E-02
NCAPD3	1.66	3.07E-02
APEH	1.46	3.07E-02
DCLK1	1.42	3.08E-02
HLF	0.74	3.09E-02
CACNA1B	0.53	3.10E-02
MICALL1	0.75	4.42E-02
USP37	0.72	4.54E-02
ACSF2	0.68	4.55E-02
UTRN	1.40	4.61E-02
C1orf192	1.55	4.69E-02
NR2C2AP	0.76	4.82E-02
FAM73A	0.70	4.91E-02

RNF5	0.60	3.10E-02
APOB48R	2.00	3.10E-02
FLJ12684	0.47	3.10E-02
GGT5	0.66	3.10E-02
SMAD6	0.70	3.10E-02
SLC38A1	1.45	3.11E-02
DCLRE1B	1.60	3.16E-02
IL3RA	0.50	3.16E-02
PDXDC2	1.34	3.16E-02
SLC39A14	1.39	3.16E-02
PRKCDBP	0.68	3.16E-02
ATP13A4	1.48	3.16E-02
CD48	2.37	3.16E-02
TRIM26	0.76	3.16E-02
ZNF219	1.74	3.20E-02
HGF	1.69	3.26E-02
CNOT6L	0.75	3.32E-02
ANO6	1.39	3.36E-02
GSPT1	0.46	3.36E-02
MIR325	0.48	3.36E-02
COL5A3	1.37	3.36E-02
ZNF326	0.68	3.38E-02
ZNF287	0.74	3.40E-02
NBPF20	0.59	3.41E-02
BBS10	0.58	3.45E-02
SLC44A3	1.35	3.46E-02
PMS2L4	0.76	3.51E-02
CDKN3	1.81	3.55E-02

TUBAL3	0.45	3.61E-02
FAM113B	1.58	3.62E-02
DNHD1	1.58	3.67E-02
EPAS1	0.77	3.67E-02
SNORA5C	0.33	3.67E-02
SBNO2	1.60	3.67E-02
NRCAM	1.36	3.72E-02
PTER	1.75	3.80E-02
RALB	0.76	3.81E-02
TRPV2	1.45	3.81E-02
FLJ20309	0.74	3.81E-02
C1orf69	0.57	3.89E-02
ADH4	0.69	3.91E-02
NCF1	1.83	3.94E-02
TTC32	1.60	3.95E-02
TSPAN6	1.51	3.96E-02
GORASP1	1.44	4.01E-02
FLJ10781	1.38	4.04E-02
NOTCH3	0.68	4.04E-02
ABCA6	0.62	4.05E-02
CROCCL2	0.68	4.10E-02
FHL5	0.60	4.11E-02
ALDH5A1	1.37	4.12E-02
EVI1	0.71	4.14E-02
HIST1H2BJ	0.61	4.14E-02
NDUFB11	1.63	4.14E-02
TMEM132E	1.87	4.18E-02
STX17	0.65	4.19E-02

HMGCR	0.63	4.21E-02
PAPSS2	0.56	4.21E-02
C12orf5	1.48	4.22E-02
MAGEF1	1.66	4.22E-02
ARHGAP19	0.70	4.24E-02
C9orf43	1.79	4.27E-02
ODZ4	1.34	4.31E-02
FHOD3	1.50	4.32E-02
OSBPL3	1.71	4.33E-02
ZSCAN2	0.64	4.38E-02
CD24	1.66	4.39E-02
DGCR6	1.56	4.39E-02
SLFN13	0.60	4.40E-02
C7orf64	0.74	4.43E-02
STAG3L1	0.76	4.44E-02
USP49	0.66	4.44E-02
GRAMD1C	1.55	4.45E-02
SYNM	1.03	4.48E-02
ZKSCAN1	0.63	4.53E-02
CCDC85A	1.37	4.53E-02
ZNF322B	0.46	4.58E-02
CD34	0.66	4.60E-02
MT1F	1.42	4.67E-02
LRRC37B2	0.74	4.67E-02
RNF138P1	0.71	4.69E-02
KIF2A	0.49	4.69E-02
UAP1	1.55	4.78E-02
HTRA4	2.10	4.78E-02

LOC283932	0.61	4.80E-02
PHF10	0.72	4.80E-02
HAMP	1.71	4.80E-02
GGH	1.36	4.80E-02
FAHD2B	1.91	4.83E-02
FBP1	1.88	4.83E-02
FCGR2B	2.17	4.83E-02
GMPR	1.45	4.86E-02
STXBP5L	1.92	4.89E-02
LYZ	1.90	4.91E-02
C10orf33	1.35	4.97E-02
LIAS	0.74	4.99E-02

Table A15. All significantly differentially expressing genes by pathological severity

Blood vessel development			Glycoprotein signalling			Glycoproteins & membrane activity			Fibronectin		
Gene	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**
SELP	0.38	1.33E-02	IGDCC4	1.67	1.42E-03	ADCY4	0.65	5.41E-03	NRCAM	1.36	3.72E-02
EMCN	0.62	2.99E-02	ADCY4	0.65	5.41E-03	SLC44A3	1.35	3.46E-02	PTPRB	0.67	3.80E-04
ACVRL1	0.63	1.51E-02	AEBP1	1.87	7.70E-05	GPR124	0.53	3.16E-04	IGDCC4	1.67	1.42E-03

LMO2	0.57	7.75E-04	NRG3	1.68	5.13E-04	HMGCR	0.63	4.21E-02	PTPRZ1	1.49	2.73E-02
EPAS1	0.77	3.67E-02	ACVRL1	0.63	1.51E-02	ADCY8	2.02	1.34E-06	F3	1.67	1.17E-02
LEPR	0.57	3.87E-03	SLC44A3	1.35	3.46E-02	TSPAN6	1.51	3.96E-02	LEPR	0.57	3.87E-03
EDN1	0.56	1.30E-02	ARSG	0.6	2.73E-02	JAG1	0.65	2.55E-03	KAL1	2.53	1.94E-05
PRRX1	1.64	1.17E-04	ADCY8	2.02	1.34E-06	IL17RB	1.8	7.12E-04	TEK	0.69	1.25E-02
JAG1	0.65	2.55E-03	HMGCR	0.63	4.21E-02	NRCAM	1.36	3.72E-02	ROBO4	0.36	1.01E-05
CD44	1.23	1.24E-02	GPR124	0.53	3.16E-04	CD48	2.37	3.16E-02	PTPRO	1.78	2.28E-02
PLCD3	1.49	2.49E-02	TRPV2	1.45	3.81E-02	LILRA2	2	1.65E-02	IL3RA	0.5	3.16E-02
ROBO4	0.36	1.01E-05	LEPR	0.57	3.87E-03	CD44	1.23	1.24E-02			
FOXC1	0.67	2.11E-02	SUSD1	1.55	1.76E-02	CH25H	2.47	4.85E-08			
SOX18	0.41	4.16E-10	TSPAN6	1.51	3.96E-02	RALB	0.76	3.81E-02			
			JAG1	0.65	2.55E-03	ELTD1	0.64	1.57E-02			
			SIRPB1	1.83	1.49E-02	DNAJC1	1.5	1.81E-02			
			IL17RB	1.8	7.12E-04	ODZ4	1.34	4.31E-02			
			NRCAM	1.36	3.72E-02	MS4A7	1.85	2.00E-02			
			CD48	2.37	3.16E-02	TMEM204	0.47	1.29E-02			
			C4ORF31	1.93	1.09E-02	VANGL1	0.46	2.82E-02			
			LILRA2	2	1.65E-02	FAM21A	0.08	9.50E-19			
			CD44	1.23	1.24E-02	PIM1	2.07	3.73E-05			
			CH25H	2.47	4.85E-08	TMEM132E	1.87	4.18E-02			
			AOAH	1.75	2.02E-02	PTPRO	1.78	2.28E-02			
			SLC1A7	0.36	9.20E-04	CD38	2.39	8.55E-06			
			ROBO4	0.36	1.01E-05	SLC25A34	0.53	4.65E-05			
			SERPINA3	1.5	2.12E-02	CD34	0.66	4.60E-02			
			CFH	0.63	1.19E-02	CHPF	1.86	6.78E-03			
			PLA1A	0.63	1.19E-03	F3	1.67	1.17E-02			
			CNTNAP1	1.52	1.99E-02	ST14	2.34	2.91E-03			
			ELTD1	0.64	1.57E-02	SLC38A1	1.45	3.11E-02			

ANO6	1.39	3.36E-02	EMCN	0.62	2.99E-02
ODZ4	1.34	4.31E-02	HLA-DRB1	2.72	1.17E-02
RAMP2	0.54	1.51E-05	FCGRT	1.59	5.23E-03
BTNL9	0.66	4.30E-03	ITGB2	1.48	1.44E-02
TMEM204	0.47	1.29E-02	NAPB	3.26	2.69E-03
OLFML2A	0.62	1.20E-02	KCNA5	1.67	2.19E-03
TMEM132E	1.87	4.18E-02	ABCA6	0.62	4.05E-02
PTPRO	1.78	2.28E-02	GORASP1	1.44	4.01E-02
PCDHGB5	0.71	2.93E-02	GPR27	0.54	1.01E-04
CD38	2.39	8.55E-06	STX17	0.65	4.19E-02
F5	2.35	2.65E-03	TEK	0.69	1.25E-02
CD34	0.66	4.60E-02	HLA-DRB4	0.09	2.56E-15
CHST6	1.56	9.20E-04	DNHD1	1.58	3.67E-02
CHPF	1.86	6.78E-03	ABCB1	0.69	3.26E-03
CHST9	2.38	8.23E-05	TRIL	1.48	1.73E-03
F3	1.67	1.17E-02	IFI27L1	1.42	6.75E-03
ST14	2.34	2.91E-03	ABCG2	0.61	2.33E-03
SLC38A1	1.45	3.11E-02	ATP13A4	1.48	3.16E-02
SLC39A14	1.39	3.16E-02	NOTCH3	0.68	4.04E-02
NAGLU	1.73	5.26E-03	GGT5	0.66	3.10E-02
EMCN	0.62	2.99E-02	EI24	0.71	3.00E-02
CCL2	1.63	2.12E-02	TNFSF10	0.35	2.92E-05
HLA-DRB1	2.72	1.17E-02	SLC4A11	1.89	1.82E-02
OXTR	0.63	7.65E-04	RNF5	0.6	3.10E-02
ITGB2	1.48	1.44E-02	LRRC37B2	0.74	4.67E-02
FCGRT	1.59	5.23E-03	C19ORF18	2.01	2.11E-04
KCNA5	1.67	2.19E-03	BEGAIN	0.46	1.48E-03
NAALADL1	0.62	1.83E-02	ABO	0.4	2.77E-04

POMC	0.58	2.42E-02	IL3RA	0.5	3.16E-02
ABCA6	0.62	4.05E-02	MS4A14	2.22	1.33E-02
DCT	1.77	1.31E-02	CACNA1B	0.53	3.10E-02
RGMA	1.72	9.04E-04	IGDCC4	1.67	1.42E-03
C1QTNF5	1.61	9.20E-04	NRG3	1.68	5.13E-04
GPR27	0.54	1.01E-04	ACVRL1	0.63	1.51E-02
KAL1	2.53	1.94E-05	TRPV2	1.45	3.81E-02
TEK	0.69	1.25E-02	LEPR	0.57	3.87E-03
HLA-DRB4	0.09	2.56E-15	CYP2S1	1.98	3.04E-02
CD24	1.66	4.39E-02	SUSD1	1.55	1.76E-02
PTPRB	0.67	3.80E-04	GJA1	1.44	3.31E-03
SELP	0.38	1.33E-02	SIRPB1	1.83	1.49E-02
IL6	2.04	1.51E-02	STXBP5L	1.92	4.89E-02
PLA2G15	1.53	1.01E-02	APOB48R	2	3.10E-02
PTPRZ1	1.49	2.73E-02	SLC1A7	0.36	9.20E-04
GGH	1.36	4.80E-02	SNTB1	1.58	9.69E-03
NTNG1	1.53	1.50E-03	CNTNAP1	1.52	1.99E-02
ABCB1	0.69	3.26E-03	ANO6	1.39	3.36E-02
HBA1	0.36	2.49E-02	TOMM34	0.67	1.01E-02
HGF	1.69	3.26E-02	RAMP2	0.54	1.51E-05
COL5A3	1.37	3.36E-02	NDUFB11	1.63	4.14E-02
TRIL	1.48	1.73E-03	BTNL9	0.66	4.30E-03
TMEM2	0.59	2.02E-04	CYYR1	0.68	1.14E-03
ABCG2	0.61	2.33E-03	PCDHGB5	0.71	2.93E-02
NOTCH3	0.68	4.04E-02	CHST6	1.56	9.20E-04
DNASE2	1.69	1.61E-04	CHST9	2.38	8.23E-05
GGT5	0.66	3.10E-02	CLIC5	0.42	1.25E-02
SLC4A11	1.89	1.82E-02	SLC41A1	1.56	8.55E-03

FCGR2B	2.17	4.83E-02	GRAMD1C	1.55	4.45E-02
ABCC3	3.11	4.39E-04	C6ORF192	1.4	1.06E-02
LRRC37B2	0.74	4.67E-02	DEGS2	0.58	7.67E-03
ABO	0.4	2.77E-04	GRASP	0.47	2.22E-03
IGFBP3	0.46	8.57E-03	SLC39A14	1.39	3.16E-02
BMP8B	0.43	2.74E-03	NKD1	0.65	2.86E-03
IL3RA	0.5	3.16E-02	OXTR	0.63	7.65E-04
CACNA1B	0.53	3.10E-02	NAALADL1	0.62	1.83E-02
GPR116	0.65	2.15E-04	GRAMD4	1.71	1.73E-03
			DCT	1.77	1.31E-02
			RGMA	1.72	9.04E-04
			C1QTNF5	1.61	9.20E-04
			DGKE	0.66	2.29E-02
			MARVELD2	0.48	3.79E-03
			PLCD3	1.49	2.49E-02
			CD24	1.66	4.39E-02
			PTPRB	0.67	3.80E-04
			SELP	0.38	1.33E-02
			PIRT	1.62	2.84E-04
			TMEM45B	0.56	2.91E-02
			PTPRZ1	1.49	2.73E-02
			NTNG1	1.53	1.50E-03
			TMEM2	0.59	2.02E-04
			RAB40AL	0.65	2.73E-02
			FCGR2B	2.17	4.83E-02
			ABCC3	3.11	4.39E-04
			CMTM8	0.58	6.07E-08
			GPR116	0.65	2.15E-04

Leukocyte migration & cell motility			Leukocyte migration & response to external stimuli			Defence response			Surface antigen & binding		
Gene	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**
SELP	0.38	1.33E-02	EDN3	0.48	1.75E-02	SELP	0.38	1.33E-02	SELP	0.38	1.33E-02
EDN3	0.48	1.75E-02	SELP	0.38	1.33E-02	IL6	2.04	1.51E-02	EMCN	0.62	2.99E-02
IL6	2.04	1.51E-02	SBNO2	1.6	3.67E-02	CCL2	1.63	2.12E-02	ACVRL1	0.63	1.51E-02
CCL2	1.63	2.12E-02	IL6	2.04	1.51E-02	NCF1	1.83	3.94E-02	ABCB1	0.69	3.26E-03
ACVRL1	0.63	1.51E-02	AOAH	1.75	2.02E-02	LYZ	1.9	4.91E-02	IL17RB	1.8	7.12E-04
HMGCR	0.63	4.21E-02	F3	1.67	1.17E-02	MYO1F	1.94	3.46E-03	NRCAM	1.36	3.72E-02
PAX6	1.82	2.39E-03	EDN1	0.56	1.30E-02	AFAP1L2	0.61	1.71E-03	CD48	2.37	3.16E-02
ITGB2	1.48	1.44E-02	CD24	1.66	4.39E-02	ITGB2	1.48	1.44E-02	CD38	2.39	8.55E-06
NRCAM	1.36	3.72E-02				IL17RB	1.8	7.12E-04	RGMA	1.72	9.04E-04
CD44	1.23	1.24E-02				MIF	1.57	7.39E-03	CD44	1.23	1.24E-02
CD34	0.66	4.60E-02				CD48	2.37	3.16E-02	CD34	0.66	4.60E-02
KAL1	2.53	1.94E-05				LILRA2	2	1.65E-02	KAL1	2.53	1.94E-05
FOXC1	0.67	2.11E-02				CD44	1.23	1.24E-02	F3	1.67	1.17E-02
CD24	1.66	4.39E-02				HAMP	1.71	4.80E-02	CD24	1.66	4.39E-02
DCLK1	1.42	3.08E-02				AOAH	1.75	2.02E-02			
						F3	1.67	1.17E-02			
						HIST1H2BJ	0.61	4.14E-02			
						SERPINA3	1.5	2.12E-02			
						CFH	0.63	1.19E-02			
						CD24	1.66	4.39E-02			

Extracellular region			Plasma membrane			Leukocyte migration & response to bacterial molecule			Phosphate activity		
Gene	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**

EDN3	0.48	1.75E-02	EMCN	0.62	2.99E-02	EDN3	0.48	1.75E-02	PTPRB	0.67	3.80E-04
AEBP1	1.87	7.70E-05	NRG3	1.68	5.13E-04	SELP	0.38	1.33E-02	PTPRZ1	1.49	2.73E-02
EMCN	0.62	2.99E-02	ACVRL1	0.63	1.51E-02	IL6	2.04	1.51E-02	PFKFB2	1.52	1.47E-02
NRG3	1.68	5.13E-04	SHROOM4	0.23	8.43E-04	CCL2	1.63	2.12E-02	FBP1	1.88	4.83E-02
CCL2	1.63	2.12E-02	HLA-DRB1	2.72	1.17E-02	CD34	0.66	4.60E-02	C12ORF5	1.48	4.22E-02
LEPR	0.57	3.87E-03	TRPV2	1.45	3.81E-02	ITGB2	1.48	1.44E-02	CDKN3	1.81	3.55E-02
EDN1	0.56	1.30E-02	OXTR	0.63	7.65E-04	CD24	1.66	4.39E-02	PTPRO	1.78	2.28E-02
JAG1	0.65	2.55E-03	GJA1	1.44	3.31E-03						
POMC	0.58	2.42E-02	FCGRT	1.59	5.23E-03						
OTOS	3.44	1.22E-02	ITGB2	1.48	1.44E-02						
MIF	1.57	7.39E-03	JAG1	0.65	2.55E-03						
IL17RB	1.8	7.12E-04	KCNA5	1.67	2.19E-03						
APOB48R	2	3.10E-02	SIRPB1	1.83	1.49E-02						
C4ORF31	1.93	1.09E-02	IL17RB	1.8	7.12E-04						
C1QTNF5	1.61	9.20E-04	CD48	2.37	3.16E-02						
CD44	1.23	1.24E-02	NRCAM	1.36	3.72E-02						
HAMP	1.71	4.80E-02	CD44	1.23	1.24E-02						
KAL1	2.53	1.94E-05	GORASP1	1.44	4.01E-02						
DGCR6	1.56	4.39E-02	MARVELD2	0.48	3.79E-03						
CFH	0.63	1.19E-02	TEK	0.69	1.25E-02						
SERPINA3	1.5	2.12E-02	SNTB1	1.58	9.69E-03						
PLA1A	0.63	1.19E-03	RALB	0.76	3.81E-02						
HTRA4	2.1	4.78E-02	HLA-DRB4	0.09	2.56E-15						
APLN	1.68	1.42E-02	CNTNAP1	1.52	1.99E-02						
SELP	0.38	1.33E-02	CD24	1.66	4.39E-02						
IL6	2.04	1.51E-02	DCLK1	1.42	3.08E-02						
PLA2G15	1.53	1.01E-02	TOMM34	0.67	1.01E-02						
PTPRZ1	1.49	2.73E-02	PTPRB	0.67	3.80E-04						

NTNG1	1.53	1.50E-03	RAMP2	0.54	1.51E-05
GGH	1.36	4.80E-02	SELP	0.38	1.33E-02
LYZ	1.9	4.91E-02	IL6	2.04	1.51E-02
OLFML2A	0.62	1.20E-02	TMEM204	0.47	1.29E-02
HGF	1.69	3.26E-02	PTPRZ1	1.49	2.73E-02
COL5A3	1.37	3.36E-02	NCF1	1.83	3.94E-02
C2ORF40	1.66	2.42E-02	NTNG1	1.53	1.50E-03
TNFSF10	0.35	2.92E-05	ABCB1	0.69	3.26E-03
CXCL14	0.53	7.39E-03	PTPRO	1.78	2.28E-02
F5	2.35	2.65E-03	RAB40AL	0.65	2.73E-02
C12ORF39	1.77	2.35E-02	NOTCH3	0.68	4.04E-02
CHST9	2.38	8.23E-05	TNFSF10	0.35	2.92E-05
F3	1.67	1.17E-02	SLC4A11	1.89	1.82E-02
ST14	2.34	2.91E-03	CD34	0.66	4.60E-02
CMTM8	0.58	6.07E-08	F3	1.67	1.17E-02
ABO	0.4	2.77E-04	ST14	2.34	2.91E-03
IGFBP3	0.46	8.57E-03	ABCC3	3.11	4.39E-04
BMP8B	0.43	2.74E-03	SYNM	1.03	4.48E-02
			GRASP	0.47	2.22E-03
			CACNA1B	0.53	3.10E-02

Leukocyte migration & chemotaxis			**EGF-like domain**		
Gene	**Fold Change**	**P-value**	**Gene**	**Fold Change**	**P-value**
EDN3	0.48	1.75E-02	DCT	1.77	1.31E-02
IL6	2.04	1.51E-02	NOTCH3	0.68	4.04E-02
CCL2	1.63	2.12E-02	SELP	0.38	1.33E-02
CXCL14	0.53	7.39E-03	NRG3	1.68	5.13E-04
HMGCR	0.63	4.21E-02	SUSD1	1.55	1.76E-02

ADCY8	2.02	1.34E-06	TEK	0.69	1.25E-02
LEPR	0.57	3.87E-03	NTNG1	1.53	1.50E-03
KAL1	2.53	1.94E-05	CNTNAP1	1.52	1.99E-02
CMTM8	0.58	6.07E-08	ELTD1	0.64	1.57E-02
OXTR	0.63	7.65E-04	ITGB2	1.48	1.44E-02
ITGB2	1.48	1.44E-02	JAG1	0.65	2.55E-03
TCF15	0.5	2.29E-02	ODZ4	1.34	4.31E-02

Table A16. List of differentially expressing genes by functional cluster for spinal regions showing mild-moderate pathology

Glycosylation		
Gene	**Fold Change**	**P-value**
SLC1A7	0.26	2.96E-05
LGI4	0.58	1.04E-04
CPA4	0.45	2.53E-03
BTN3A2	1.64	7.31E-03
GPR107	0.46	7.31E-03
GPR27	0.47	8.46E-03
RAMP2	0.65	1.28E-02
CFH	0.57	2.32E-02
FFAR3	0.48	2.96E-02
ADCY4	0.59	3.18E-02
HLA-C	2.51	3.18E-02
HLA-H	0.60	4.88E-02

Table A17. List of differentially expressing genes by functional cluster for spinal regions showing moderate pathology

Metabolic and biosynthetic processes		
Gene	**Fold Change**	**P-value**
GGT5	0.71	3.51E-02
MSRA	1.56	1.38E-03
MTHFD2L	0.74	2.78E-02
GSTT1	2.55	3.90E-02
SLC19A3	0.67	1.16E-02
PAPSS2	0.61	4.21E-02

Table A18. List of differentially expressing genes by functional cluster for spinal regions showing moderate-severe pathology

Striated muscle cell development			Microsomes and vesicular fraction			Fatty acid metabolic processes		
Gene	Fold Change	P-value	Gene	Fold Change	P-value	Gene	Fold Change	P-value
UTRN	1.40	4.61E-02	DCT	1.92	2.62E-03	ACSL1	0.67	9.33E-03
MYOZ1	1.35	3.82E-02	FMO4	1.82	1.25E-02	PTGS2	0.40	3.14E-05
CHRNA1	2.90	3.56E-03	CYP3A5	1.74	1.34E-03	ALDH5A1	1.39	2.35E-02
SNTA1	1.45	9.09E-03	PPP1R3C	0.76	2.89E-02	CYP4F3	1.98	2.23E-02
			LRAT	1.38	9.09E-03	ACACB	1.39	1.91E-02
			ACSL1	0.67	9.33E-03	ACSF2	0.68	4.55E-02
			PTGS2	0.40	3.14E-05	ACSM5	2.56	1.18E-03
			HMGCR	0.66	1.72E-02			
			CYP4F3	1.98	2.23E-02			
			CALR	0.65	3.83E-02			
			FOSL1	2.10	1.94E-02			

Table A19. List of differentially expressing genes by functional cluster for spinal regions showing severe pathology

Probe	Gene	P-value	Adj. P	F Statistic
6510307	ST5	3.88E-08	7.78E-04	46.68845
4260424	PARD3B	2.18E-07	2.19E-03	36.52398
510669	CNTFR	3.75E-06	2.18E-02	23.90182
2750139	RNF216L	4.34E-06	2.18E-02	23.36336
1190082	C6ORF48	2.24E-05	8.96E-02	17.98008
3940309	CDAN1	4.23E-05	1.41E-01	16.16605
7610189	ZNF706	6.03E-05	1.73E-01	15.21642
4150537	ZNF331	8.38E-05	2.00E-01	14.37108
540706	HLF	0.000101	2.00E-01	13.91797
4230021	NCK2	0.000108	2.00E-01	13.75316
7000736	MLC1	0.00011	2.00E-01	13.70059
1570687	IFT122	0.000127	2.03E-01	13.35793
670326	ABI2	0.000132	2.03E-01	13.26287
380224	FADS2	0.000153	2.13E-01	12.91631
4900670	COX5A	0.00016	2.13E-01	12.81235
460577	STON2	0.000178	2.23E-01	12.55788
430546	HIST1H2BG	0.000266	3.07E-01	11.66773
1570767	HOMER2	0.000311	3.07E-01	11.32783
7150349	TPD52L1	0.000329	3.07E-01	11.21263
3370136	ACAD11	0.000338	3.07E-01	11.1518
5550500	MXRA7	0.000355	3.07E-01	11.05158
3130669	SATB1	0.000357	3.07E-01	11.03642
5820681	PBX2	0.000358	3.07E-01	11.02964
2760678	RCVRN	0.000406	3.07E-01	10.76943
2320750	MRVI1	0.000414	3.07E-01	10.7303
6770131	OLFM4	0.000419	3.07E-01	10.70203
7560136	WDR92	0.000441	3.07E-01	10.59726

6940348	TIPIN	0.000456	3.07E-01	10.5314
6330168	EVC	0.000511	3.07E-01	10.30169
3290400	CUTA	0.000528	3.07E-01	10.23488
3520750	SPOCK3	0.000569	3.07E-01	10.08545
1340685	PLEC1	0.000572	3.07E-01	10.07721
6060358	ATG4A	0.000593	3.07E-01	10.0031
6560167	DIAPH3	0.000594	3.07E-01	10.00103
3930008	RPL14	0.000603	3.07E-01	9.972876
4480767	NUP98	0.000611	3.07E-01	9.945566
6110095	CCDC15	0.000617	3.07E-01	9.926901
3360291	TICAM1	0.000649	3.07E-01	9.826936
4220731	P4HA1	0.000681	3.07E-01	9.732617
3310167	SDF4	0.000712	3.07E-01	9.647009
1740471	TINF2	0.000717	3.07E-01	9.633339
5340414	LGTN	0.000727	3.07E-01	9.606422
6840494	CYP4F3	0.000752	3.07E-01	9.542619
5130551	SOX6	0.000779	3.07E-01	9.47462
4610136	GPATCH4	0.00079	3.07E-01	9.447991
6330133	BPGM	0.000794	3.07E-01	9.438416
6590731	DTNBP1	0.000807	3.07E-01	9.40863
2900593	VEGFA	0.000821	3.07E-01	9.374119
3520196	LIG4	0.000828	3.07E-01	9.35835
3930603	RPL23AP53	0.000845	3.07E-01	9.320729
4050368	AGFG2	0.00087	3.07E-01	9.265605
6580524	KIAA1161	0.000879	3.07E-01	9.246611
130669	BAG5	0.000883	3.07E-01	9.238592
6510093	PPARA	0.0009	3.07E-01	9.202857
7380167	C21ORF69	0.000901	3.07E-01	9.20069

3830682	TTC9C	0.000901	3.07E-01	9.199352
3990070	PCDHGA5	0.000909	3.07E-01	9.182836
270204	RFNG	0.000913	3.07E-01	9.175077
4050593	MRC2	0.000924	3.07E-01	9.152795
5560474	RAMP2	0.00094	3.07E-01	9.121292
4200093	MRO	0.000944	3.07E-01	9.113269
7650070	RAMP2	0.000955	3.07E-01	9.090783
4920747	ZNF823	0.000966	3.07E-01	9.069866

Table A20. Genome-wide One-Way ANOVA of gene expression changes by severity level

	EQ-5D Utility		**VAS Scores**	
	Missing Value Frequency	**Missing Value %**	**Missing Value Frequency**	**Missing Value %**
Stage 2a	139	65%	139	65%
Stage 2b	89	41·6%	89	41·6%
Stage 3	71	33·2%	72	33·6%
Stage 4	147	68·7%	147	68·7%

Table A21. Missing values for EQ-5D Utility and VAS scores 214 LiCALS patients

	Mobility			***Change* in Mob when moving to ALS stage**								
				2b			**3**			**4**		
ALS stage	**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**	
2a	1·63	1·54	1·73	0·16**	0·07	0·25	0·37**	0·27	0·46	0·53**	0·42	0·63
2b	1·80	1·72	1·88				0·21**	0·13	0·28	0·36**	0·27	0·45
3	2·00	1·93	2·08							0·16**	0·07	0·24
4	2·16	2·07	0·25									
	Self-Care			***Change* in SC when moving to ALS stage**								
				2b			**3**			**4**		
ALS stage	**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**	
2a	1·65	1·53	1·77	0·25**	0·14	0·37	0·50**	0·38	0·63	0·77**	0·63	0·91
2b	1·91	1·80	2·01				0·25**	0·15	0·35	0·52**	0·40	0·63
3	2·16	2·06	2·25							0·26**	0·15	0·37
4	2·42	2·31	2·54									
	Usual Activity			***Change* in UA when moving to ALS stage**								
				2b			**3**			**4**		
ALS stage	**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**	
2a	1·90	1·79	2·01	0·20**	0·09	0·31	0·36**	0·25	0·47	0·60**	0·47	0·73
2b	2·09	2·00	2·16				0·16**	0·07	0·25	0·40**	0·29	0·51

3	2·26	2·17	2·34							0·24**	0·14	0·34
4	2·50	2·39	2·60									
	Pain/Discomfort			***Change* in PD when moving to ALS stage**								
				2b			**3**			**4**		
ALS stage	**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**	
2a	1·37	1·26	1·47	0·18**	0·07	0·29	0·24**	0·13	0·35	0·20**	0·07	0·33
2b	1·55	1·46	1·63				0·06**	-0·03	0·15	0·02**	-0·08	0·13
3	1·60	1·53	1·68							-0·04**	-0·13	0·06
4	1·57	1·47	1·67									
	Anxiety/Depression			***Change* in AD when moving to ALS stage**								
				2b			**3**			**4**		
ALS stage	**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**		**Mean**	**95% CI**	
2a	1·32	1·22	1·42	0·13**	0·02	0·23	0·11**	0·00	0·22	0·24**	0·12	0·36
2b	1·45	1·37	1·53				-0·02**	-0·01	-0·07	0·11**	0·00	0·21
3	1·43	1·36	1·50							0·13**	0·03	0·22
4	1·56	1·46	1·65									

Table A22. Comparing differences in EQ-5D dimension scores across ALS clinical stage using MLE regression
* $p < ·05$
** $p < ·01$

Figures Appendix

Chapter 2

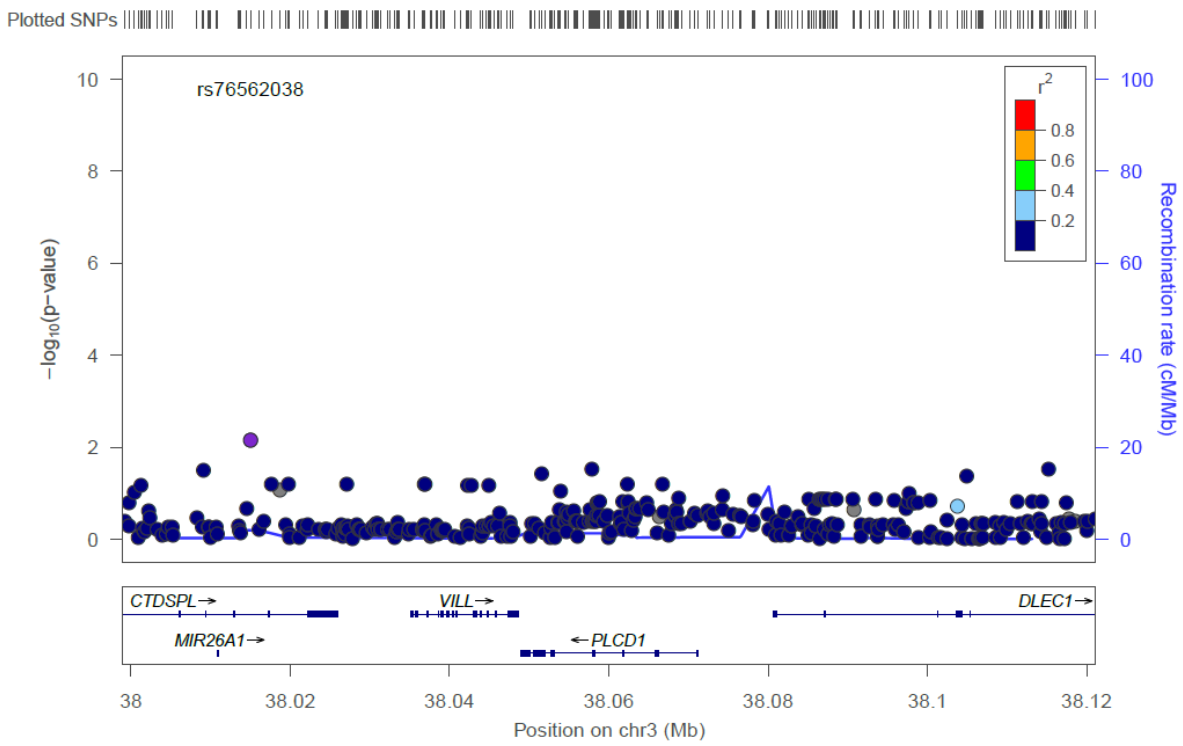


Figure A1. LocusZoom Manhattan plot of imputed SNPs in and around PLCD1 for the international dataset. The heat scale represents linkage disequilibrium. Bonferroni-correction alpha value was at 2.82×10^{-5}.

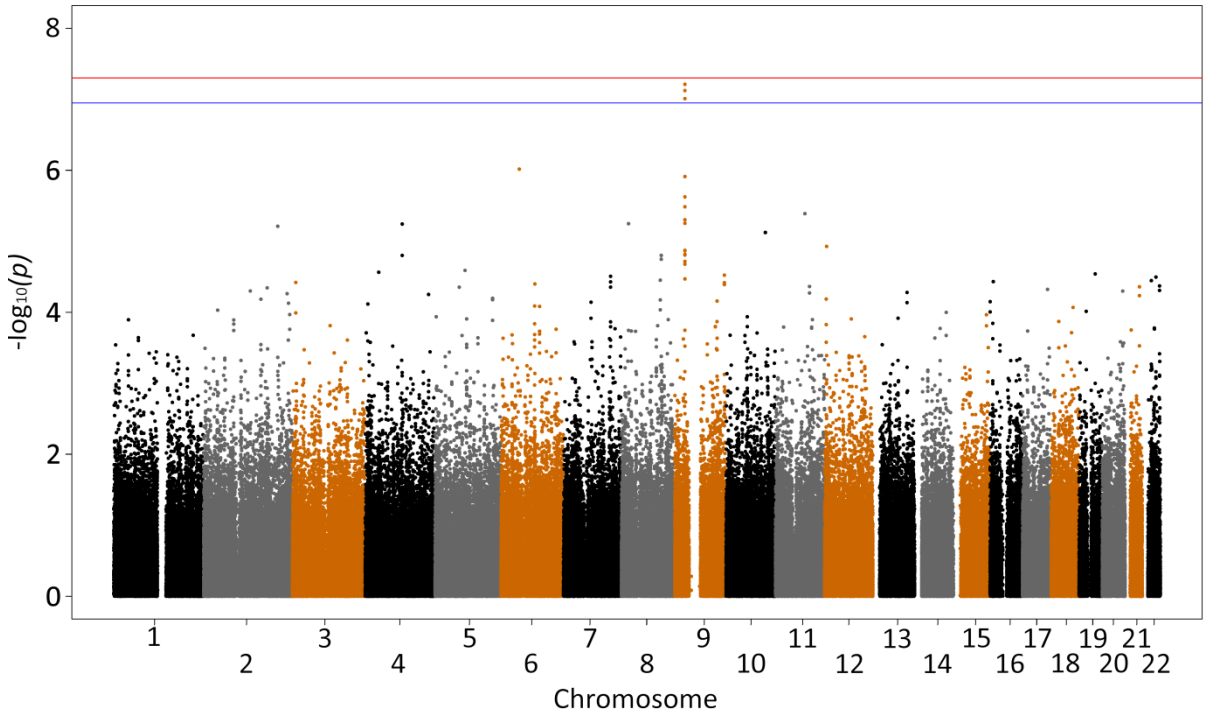


Figure A2: GWAS of All Cases (599) vs. Controls (4136). Red line is genome correction threshold (5 x 10^{-8}) and blue line is bonferroni threshold (1.13 x 10^{-7}).

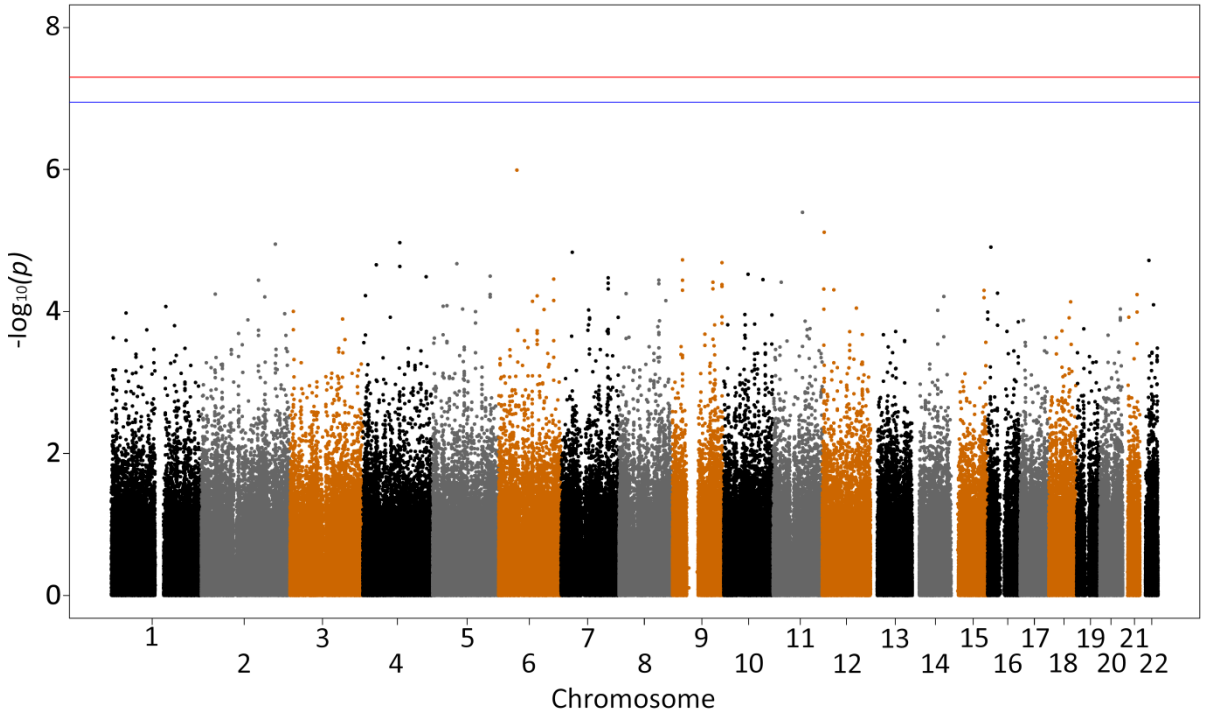


Figure A3. GWAS of mutation cases removed (560) vs. controls (4136). Red line is genome correction threshold (5 x 10-8) and blue line is bonferroni threshold (1.13 x 10^{-7}).

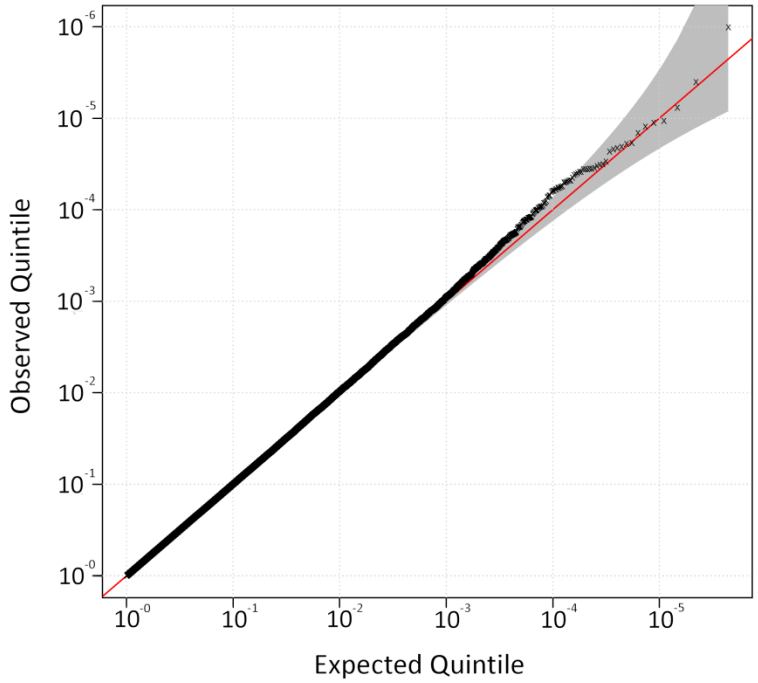


Figure A4. Q-Q Plot for GWAS comparing non-repeat mutation cases (560) vs. controls (4136)

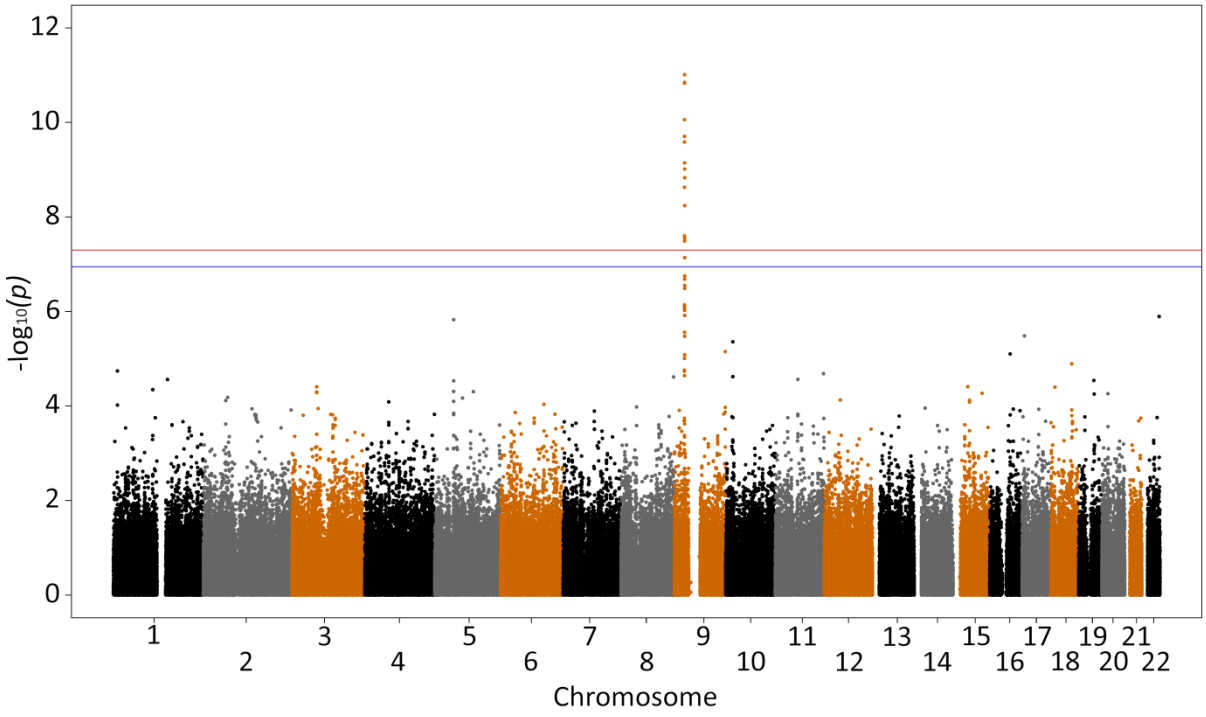


Figure A5. Manhattan plot of GWAS comparing repeat mutation cases (39) vs. controls (4136). Red line is genome correction threshold (5 x 10^{-8}) and blue line is bonferroni threshold (1.13 x 10^{-7}).

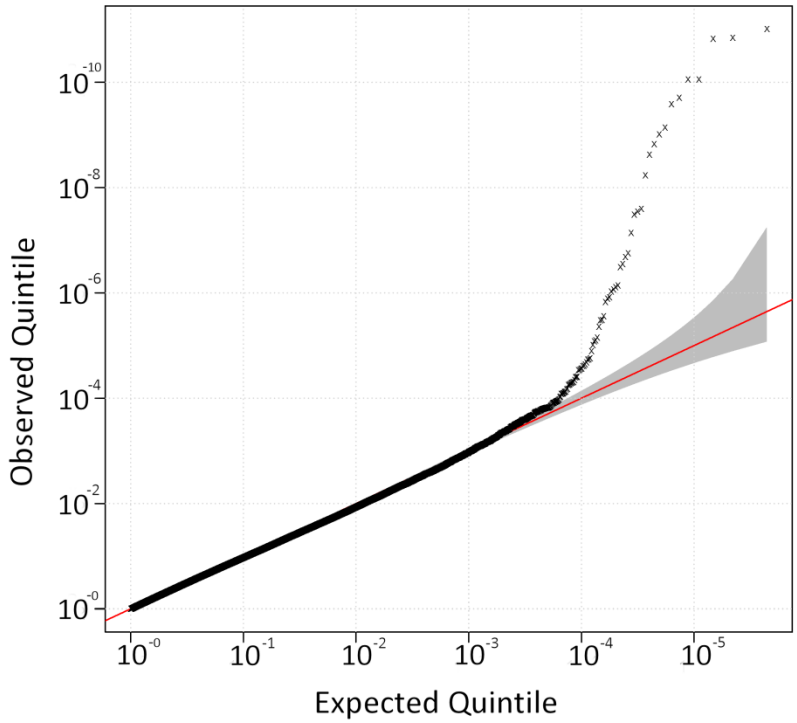
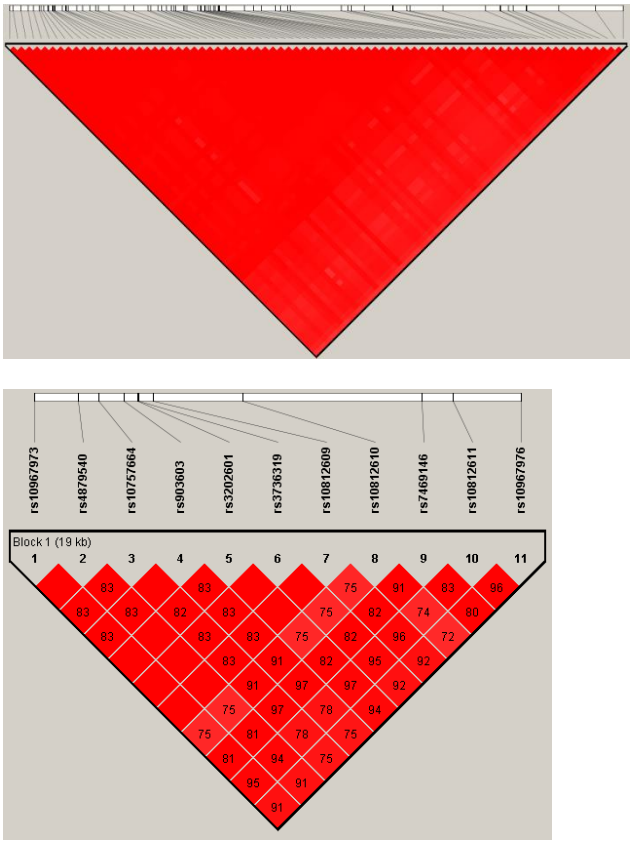


Figure A6. Q-Q Plot for GWAS comparing repeat mutation cases (39) vs. controls (4136) (λ_{GC} of 1.02).



79-SNP Mutation-Specific Haplotype	**11-SNP Residual Haplotype**
Haplotype association analysis of mutation cases (39) vs. controls (1182): $\chi^2 = 64.62$; $p = 9.08 \times 10^{-14}$	Haplotype association analysis of non-mutation cases (560) vs. controls (1182): $\chi^2 = 15.58$; $p = 7.90 \times 10^{-5}$

Figure A7. 79-SNP haplotype for cases with the pathological expansion (left) and an 11-SNP haplotype for those without (right).

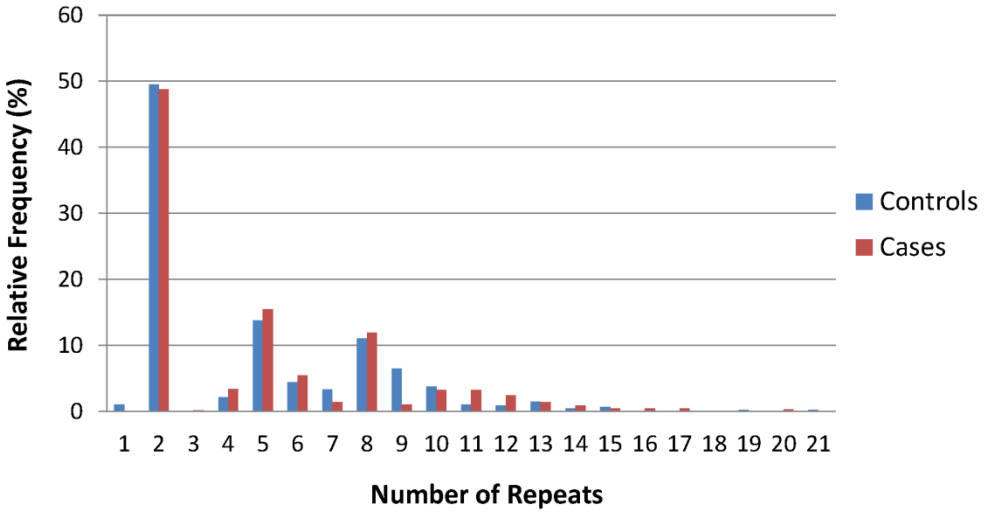


Figure A8. Relative frequency of non-mutation cases and controls by the largest repeat length

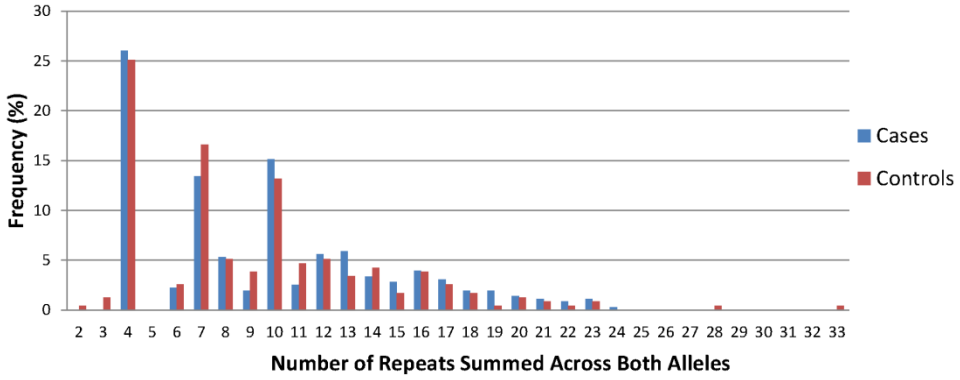


Figure A9. Relative frequency of non-repeat mutation cases (n = 239) and controls (n = 357) by repeat length as an additive model

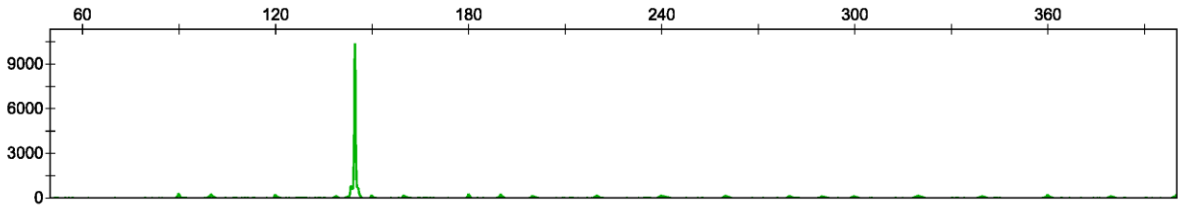


Figure A10. AFLP of example case showing homozygosity for five repeats

Data Analysis Appendix

Chapter 3

The Kruskal-Wallis test was used to examine the effect of mutation status on site of symptom onset. Age of symptom onset and the effect of onset location on age of onset were examined under one-way ANOVA, with Welch's F correction for lack of homogeneity.

Clinical information was available for 581 cases (Table S1). There was no effect of mutation status on site of onset or on age of onset. There was a significant effect of site of onset by age of onset ($F = 8.642$ (3), $p < 0.001$), consistent with previous studies.[S1.]

Data Analysis D1. Patient Characteristic Analyses

Scripts Appendix

Chapter 2 & 3

```
--bfile example_file
        --maf 0.002
        --hwe 0.00001
        --geno 0.05
        --ci 0.95
        --sex
        --adjust
        --make-bed
        --out  example_file_QC
        --noweb
```

Script A1. Typical Plink script for quality control.

This protocol is based on the ENIGMA imputation protocol http://enigma.loni.ucla.edu/protocols/genetics-protocols/) and has been modified by Dr Caroline Johnston, a bioinformatician at the Institute of Psychiatry, and myself.

Step 1. Download extra files for crossover:

```
wget "http://enigma.loni.ucla.edu/wp-content/uploads/2012/04/v2.20101123.ENIGMA2.EUR.autosomes.vcf.tgz"

wget "http://enigma.loni.ucla.edu/wp-content/uploads/2012/04/v2.20101123.ENIGMA2.EUR.autosomes.extras.tgz"

tar -xvzf v2.20101123.ENIGMA2.EUR.autosomes.vcf.tgz

tar -xvzf v2.20101123.ENIGMA2.EUR.autosomes.extras.tgz

gzip -d 1kgp.*.gz
```

Step 2. Data Quality Control

This steps removes strand ambiguous SNPs, controls for low minor allele frequency, missingness, HWE, and SNP duplications.

```
awk '{ if (($5=="T" && $6=="A")||($5=="A" && $6=="T")||($5=="C" && $6=="G")||($5=="G" && $6=="C")) print $2, "ambig" ; else print $2 ;}' Test.bim | grep ambig > ambig.list
```

plink --bfile Test --exclude ambig.list --make-founders --out lastQC --maf 0.01 --geno 0.05 --mind 0.05 --hwe 0.000001 --make-bed --noweb

plink --bfile lastQC --exclude duplicated.list --make-bed --noweb --make-founders --out lastQC2 --maf 0.01 --geno 0.05 --mind 0.05 --hwe 0.000001

Step 3. Shift pedigree file to NCBI build 37 and flip strand

This step updates the pedigree file to NCBI build 37 and flips strand ready for the 1000 genomes reference panel

awk '{print $2,$1,$3,$4,$5,$6}' lastQC.bim > tempQC.bim

awk 'NR==FNR{s=$1;a[s]=$0;next} a[$1]{print $0 " "a[$1]}' tempQC.bim 1kgp.alleles > merged.alleles

awk '{ if ($2!=$8 && $2!=$9) print $1}' merged.alleles > flip.list

plink --bfile lastQC --extract 1kgp.snps --update-map 1kgp.chr --update-chr --flip flip.list --make-bed --out temp --noweb

plink --bfile temp --update-map 1kgp.bp --make-bed --out lastQCb37 --noweb

Step 4. Calculating how many SNPs have been lost by the leftover

wc -l lastQCb37.bim

475120 lastQCb37.bim

wc -l Test.bim

488434 Test.bim

Step 5. Reformat the dat files

Recode QC'd pedigree file by –recode.

awk '{ print "M", $1 ":" $4}' lastQCb37.map > ready4mach.dat

I will replace the first column with the SNPs from original map file; this will mean the output will correspond with rs numbers rather than a Variant Call Format (VCF).

To remove the phenotype column (incompatible with MACH and minimac imputation):

awk '{$6=""; print $0}' lastQCb37.ped | sed 's/\ \ /\ /' > ready4mach.ped

Step 6. Phasing

mach1 -d lastQCb37.dat.gz –p lastQCb37.ped.gz –prefix lastQCb37 –rounds 20 –states –phase –sample 5 > lastQCb37.mach.log

Step 7. Imputation using MACH (good for GenGen and Plink) or minimac (faster imputation, mach2dat and ProbABEL)

Using MACH:

awk '{ if ($1 == "M") print $2; }' < lastQCb37.dat > lastQCb37.snps

mach1 -d lastQCb37.dat -p lastQCb37.ped -s lastQCb37.snps -h lastQCb37.hap --crossover lastQCb37.rec --errormap lastQCb37.erate --autoFlip --greedy --mle --mldetails --prefix imputation > imputation.log

Using Minimac:

minimac --vcfReference --rounds 5 --states 200 --refHaps chr9.phase1_release_v3.20101123.snps_indels_svs.genotypes.refpanel.ALL.vcf –haps lastQCb37 –snps lastQCb37.snps –autoClip lastQCb37.dat –prefix lastQCb37_imputed > lastQCb37_minimac.log

Optional Step 8: GenGen

sed -n '/rs_example_SNP1/,/ rs_example_SNP2/p' chr_example.legend > chr_example_out.legend

/.../convert_mach.pl lastQCb37.mlgeno lastQCb37.mlinfo lastQCb37 -legend chr_example_out.legend --chr example -ped ready4mach.ped -prefix imputed_gengen

Update disease status using Plink, as it would have been set to none (-9).

Script A2. Imputation and GenGen protocol using 1000 genomes.

Chapter 3

Step 1. Create Convertf parameter file

genotypename: example.ped

snpname: example.map

indivname: example.ped

outputformat: EIGENSTRAT

genotypeoutname: example.eigenstratgeno

snpoutname: example.snp

indivoutname: example.ind

familynames: NO

Step 2. Convert Pedigree files for Eigenstrat

$parfile="par.PED.EIGENSTRAT";

system("/.../eigensoft_3.0/bin/convertf -p par.EIGENSTRAT.PED")

Step 3. SmartPCA script and run

genotypename: example.eigenstratgeno

snpname: example.snp

indivname: example.ind

evecoutname: example.pca.evec

evaloutname: example.eval

altnormstyle: NO

numoutevec: 10

numoutlieriter: 5

numoutlierevec: 10

outliersigmathresh: 6.0

qtmode: 0

Step 4. Eigenstrat

/.../bin/eigenstrat -i example.eigenstratgeno -j example.pheno -p example.pca -l 10 –o eigenstrat_out

Script A3. Smart PCA and Eigenstrat scripts for identifying sub-population structure

```
--bfile example_file
        --covariates eigenstrat_out
        --maf 0.002
        --hwe 0.00001
        --geno 0.05
        --ci 0.95
        --sex
        --adjust
        --logistic
        --out  example_file_logistic
        --noweb
```

Script A4. Example Plink analysis comparing allele frequency between cases and controls using logistic regression, with quality control measures.

```
plink --file example --hap-assoc  --hap-window
2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,3
5,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50 --out windows_out --noweb
```

```
plink plink –bfile example -hap-snps
rs2239194,rs3184504,rs739496,rs10849949,rs2073950,rs2301621,rs10774625,rs6490162,rs6
28825,rs630512,rs16941541,rs616513,rs12369009 --chap --out chap_out --noweb
```

Script A5. Example Plink analysis for sliding window analyses and haplotype association test

Chapter 4

Convert Illumina to Sanger

maq ill2sanger INPUT.illumina.fastq OUTPUT.sanger.fastq

Alignment to reference transcriptome using TopHat

/.../bin/tophat -F 0.10 -a 5 -p 1 -o /.../output/ sequence.fastq

Sort bam and index file

/.../bin/samtools sort accepted_hits.bam sorted.bam

/.../samtools index sorted.bam

Add header to bam file, convert to SAM file, and isolate region of interest

/.../samtools view -b –S -h –o chr8_sort.bam sort.bam chr8:27,502,505-27,533,066

Identify SNPs and Indels using SAMtools, bcf and vcf tools, or GATK

/.../samtools mpileup -ugf /.../bowtie/indexes/hg18.fa chr8_sort.bam > rawhits.bcf

/.../bcftools view -vcg rawhits.bcf | /.../bcftools view rawhits.bcf | /.../vcfutils.pl varFilter -D100 > sort_chr8.vcf

Or

java -Xmx4g -jar /.../GenomeAnalysisTK.jar -R /.../bowtie/indexes/hg18.fa fa -T UnifiedGenotyper -I chr8_sort.bam > GATK.vcf -D /.../dbsnp_129_hg18.rod -stand_call_conf 30.0 -stand_emit_conf 10.0 -dcov 50 --platform SOLEXA

Estimate and test for differential expression comparing it to reference transcriptome

cufflinks -b /.../bowtie/indexes/hg18.fa chr8_sort.bam -o /.../cufflinks/

cuffcompare transcripts.gtf -r reference_genome/hg18.UCSCknowngene.isoforms.gtf -o /.../cuffcompare/

Using MISO to identify alternative transcripts

python run_miso.py --summarize-samples SE/control/ SE/control/

Using SpliceMap to identify novel spice junctions

/.../runSpliceMap /.../run.cfg

Run.cfg is a configuration file and uses the parameters below.

```
##
########################################################
#
# This cofiguration file contains all settings for a run
# of SpliceMap.
#
# lines begining with '#' are comments
# lists begin with '> tag' and end with '<' on separate lines
#
########################################################
##

############################
## Required Settings
##



##
# Directory of the chromosome files in FASTA format
# Each chromosome should be in a separate file (can be concatenated)
# ie. chr1.fa, chr2.fa, ...
# (single value)

genome_dir = /scratch/project/groupammar/Ash/SpliceMap/ChromosomesHG18/

##
# These are the two lists of sequencer reads files.
# "reads_list2" can be commeted out if reads are not paired-end.
# Make sure the order of both lists are the same!
# Also, "reads_list1" must be the first pair.
```

```
# Note: pair-reads should be in the "forward-reverse" format.
# (multiple values)

> reads_list1
/scratch/project/groupammar/Ash/Controls/s_1_sequence_C_BA9_2.bam/chr8.sorthead.fastq
<

#> reads_list2
#/scratch/project/groupammar/Ash/SpliceMapExample/data/long_reads_2_100K.txt.seq
#<



##
# Format of the sequencer reads, also make sure reads are
# not split over multiple lines.
# Choices are: FASTA, FASTQ, RAW
# (single value)

read_format = FASTQ

##
# Format of the quality string if FASTQ is used
# Choices are:
#   phred-33 -- Phred base 33 (same as Sanger format) [default]
#   phred-64 -- Phred base 64 (same as Illumina 1.3+)
#   solexa   -- Format used by solexa machines
# (single value)


#quality_format = phred-33



##
# The short reads mapper used
```

```
# choices are "bowtie", "eland" or "seqmap"
# Bowtie is recommended
# this can be commented out if the mapping has already been done and the
# appropriate ".t" files exist in the temp directory [advanced].
# (single value)

mapper = bowtie



##########################
## Optional Settings
##



##
# these are annotations used to find novel junctions, in ref or bed format.
# (optional)
# (single value)

annotations = /scratch/project/groupammar/Ash/SpliceMapExample/all.gene.refFlat.txt

##
# Directory name of the directory that stores temporary files
# (optional) Default = temp
# (single value)

temp_path = /scratch/project/groupammar/Ash/SpliceMap/Temp/s_1_sequence_C_BA9_2.bam/



##
# Directory name of the directory that stores the output files
# (optional) Default = output
# (single value)
```

out_path = /scratch/project/groupammar/Ash/SpliceMap/Output/s_1_sequence_C_BA9_2.bam/

##

Maximum intron size, this is absolute 99th-percentile maximum.

Introns beyond this size will be ignored.

(optional) If you don't set this, we will assume a mamalian genome (400,000)

(single value)

max_intron = 400000

##

25-th intron size, this is the lower 25th-percentile intron size

This is not the smallest size that SpliceMap will search. That is about ~25bp.

(optional) If you don't set this, we will assume a mammalian genome (20,000)

(single value)

min_intron = 20000

##

Maximum number of multi-hits

If a 25-mer seed has more than this many multi-hits, it will be discarded.

(optional) Default is 10

(single value)

max_multi_hit = 10

##

Full read length

SpliceMap will only use the first "full_read_length" bp for mapping.

If the read is shorter than "full_read_length", the full read will be used before head clip.

If you comment this out SpliceMap will use as many as possible.

This is for the case where the reads might have N's at the end.

It is always desireable to cut off the N's

(optional)

(single value)

full_read_length = 70

##

Number of bases to clip off the head of the read

This clipping is applied after "full_read_length"

(optional)

(single value)

head_clip_length = 0

##

Number of mismatches allowed in half-seeding

Choices are 0,1(default) or 2

(optional)

(single value)

seed_mismatch = 1

##

```
# Maximum number of mismatches allowed in entire read
# No limit on value, however SpliceMap can only identify reads with
# a maximum of 2 mismatches per 25bp.
# Default is 2.
# (optional)
# (single value)

read_mismatch = 2



##
# Maximum number of bases allowed to be soft clipped from the ends of reads during
# alignment. This is required as mismatches near junctions could cause parts of a
# a read to not map.
# Default is 40.
# (optional)
# (single value)

#max_clip_allowed = 40




##
# Generate a SAM file
# choices are
#   "cuff" -- Generate Cufflinks compatible SAM file
#   "sam"  -- Generate regular SAM file
#
# If this is commented out, no SAM file will be generated
# (optional)
# (single value)

sam_file = cuff
```

##

Generate separate coverage for non-uniquely mappable reads

choices are

"yes" -- Generate it [This requires 15G+ of memory for 23 million reads]

"no" -- don't generate it [default]

(optional)

(single value)

ud_coverage = yes

##

Name of the chromosome file with wildcards

If none is given the default of "chr*.fa" will be used [this is compatible with the UCSC genome files]

If your genome file is concatenated just type the file name here

(optional)

(single value)

#chromosome_wildcard = chr*.fa

##

Number of chromosomes to process at once, to take advantage of multi-core systems.

This is not threading, so it will take extra memory. However, running 2 at a time should be fine.

(optional) Default = 1

(single value)

num_chromosome_together = 2

```
##
#####################################################
#
#  Bowtie specific options
#   these have no meaning if another mapper is used
#
#####################################################
##

#########################
## Required Settings
##



##
# Base of bowtie index, this should be the same genome as the
# chromosome files
# eg. if you bowtie files are "genome/hg18/genome.1.ewbt", É
# then your base dir is "genome/hg18/genome"
# (single value)

bowtie_base_dir = /project/groupammar/RNAseq_Ammar/Ash/reference_genome/




#########################
## Optional Settings
##

##
# Number of threads to use for mapping (SpliceMap may use this option in future)
# Default value is 2
# (optional)
# (single value)
```

```
num_threads = 2



##
# Try hard?
# choices are "yes" or "no"
# Default value is yes. (it is not much slower, about 15%)
# I'm unsure if this is required, feel free to try with
# this option off and let me know your results.
# (optional)
# (single value)
try_hard = yes
```

Script A6. Scripts used for RNA-seq analyses.

Protocol Appendix

Chapter 3

Touchdown PCR protocol

95°C	3min	
94°C	1min	
75°C	1min; decreasing temperature each cycle - 1°C"	Cycle x 25 times
68°C	2min	

94°C	30sec	
56°C	30sec	
72°C	1min	Cycle x 25 times
72°C	5min	
-10°C	END	

Protocol P1. Rutherford NJB, DeJesus- Hernandez MBS, Baker MCB, Kryston TBMS, Brown PEMS, Lomen-Hoerth CMDP, et al. C9ORF72 hexanucleotide repeat expansions in patients with ALS from the Coriell Cell Repository. Neurology. 2012;79(5):482-3.